

**A CORPUS-DRIVEN STUDY ON TRANSLATION UNITS IN
AN ENGLISH-CHINESE PARALLEL CORPUS**

by

Weiqun WANG

**A thesis submitted to the School of Humanities of the University of Birmingham for the
degree of MPhil (B) in Corpus Linguistics**

**School of Humanities
The University of Birmingham**

2006

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

It is widely known that texts are not translated word by word but in larger units which are, from the perspective of the target language, more or less monosemous. This dissertation argues that translation units are the smallest such units, and that they can be identified in parallel corpora. It aims to show that these translation units and their target language equivalents can be extracted from parallel corpora and can be re-used to facilitate new translations. The concept of translation units and their equivalents will enable translators to translate competently into languages other than their native language, something not sufficiently supported by traditional bilingual dictionaries. For my exploratory study presented here, I will use the Hong Kong Legal Document Parallel Corpus (HKLDC). This dissertation starts with the definition of the concept of the translation unit and its equivalent and goes on to describe a method of extracting translation unit candidates. These candidates are then validated by further analysis. It will also test the hypothesis that each complete translation unit has only one translation equivalent. Finally, by comparing the translation equivalents extracted from the corpus with those provided by traditional dictionaries, this dissertation will argue that parallel corpora, as the repository of the translation units and translation equivalents, can, by complementing traditional translation aids, facilitate translation.

Acknowledgements

I am very grateful to Prof. Wolfgang Teubert for his thoughtful, patient and constructive supervision. I have learned a great deal from him. His profound knowledge and strict attitude will guide me forever. I will always remember his advice that because I am a non-native speaker of English, I will have to write my thesis fifteen times if a native English speaker needs to write it five times.

I am indebted to Prof. Susan Hunston as my teacher and academic adviser for her comments and patient guidance. I also would like to thank Dr. Pernilla Danielsson for her help in many ways, including providing me with the first Perl programme which enabled me to extract all the examples of Adjective plus Noun.

Special thanks should go to Paul Simmonds, for his constant encouragement and proofreading. I also would like to thank Dr. Chris Games, Dorothy Vuong and Dominc Smith who have given their friendship and understanding in reading parts of my thesis at different stages of writing.

I would, in particular, like to give special thanks to my husband Jun and my daughter Julia for their love and support during my studies. I also want to say thank you to all members of my family: my parents, brothers and sisters, and in-laws for their affection and support.

Dedicated To

Jun He

Julia He

Mrs Jianhui Li (my Mother-in-Law)

Table of Contents

TABLE OF CONTENTS.....	I
INDEX OF TABLES, EXAMPLES, FIGURES AND PICTURES.....	III
CHAPTER 1	
INTRODUCTION	1
1.1 BACKGROUND	2
1.2 AIM	5
1.3 ORGANIZATION	7
CHAPTER 2	
TRANSLATION UNIT: FROM THEORY TO PRACTICE	9
2.1 THE UNIT OF MEANING IN CORPUS LINGUISTICS.....	9
2.1.1 <i>Word – Traditional unit of meaning</i>	10
2.1.2 <i>The Foundation of Changing the View—Firth’s Theory about Meaning</i>	11
2.1.3 <i>Lexical Item — the Extended Unit of Meaning</i>	13
2.2 TRANSLATION UNIT IN CORPUS LINGUISTICS.....	16
2.2.1 <i>Translation Unit and Translation Equivalent</i>	16
2.2.2 <i>The Relationship of Translation Unit and Unit of Meaning</i>	19
2.2.3 <i>Criteria of Identifying Translation Unit in this Paper</i>	20
2.3. TRANSLATION UNIT AND PARALLEL CORPORA	22
2.3.1 <i>Types of Parallel Corpora</i>	22
2.3.2 <i>The Translation Unit Study vs the Corpus-based Translation Studies</i>	23
2.4. SUMMARY.....	24
CHAPTER 3	
THE HONG KONG LEGAL DOCUMENT CORPUS (HKLDC)	26
3.1 TEXTS	27
3.2 ANNOTATION	30
3.3 ADVANTAGES AND SHORTCOMINGS	31
3.4 SUMMARY.....	32
CHAPTER 4	
METHODOLOGY OF EXTRACTION TRANSLATION UNITS AND THEIR EQUIVALENTS	34
4.1 TRANSLATION UNITS AND A+N PATTERN	35
4.2 IMPLEMENTATION OF EXTRACTION TRANSLATION UNIT CANDIDATES AND THEIR TRANSLATION EQUIVALENTS	36
4.3 DISCUSSION ON THE COMPLETE EXTRACTION.....	40
4.4 SUMMARY.....	41
CHAPTER 5	
TRANSLATION EQUIVALENTS OF A+N PHRASES	43
5.1. THE OVERVIEW OF A+ N TRANSLATION EQUIVALENTS	43
5.1.1 <i>The Occurrence of the Phrases in the Extracted Sentences</i>	43
5.1.2 <i>The Frequency Calculation of the Phrases and Their Translation Equivalents</i>	46
5.1.3 <i>The Profile of the Chinese Translation Equivalence</i>	48
5.2 THE A+N PHRASES WITH ONE TRANSLATION EQUIVALENT	50
5.2.1 <i>A+N phrases Functioning as Translation Units</i>	50
5.2.2 <i>The A+N Phrases Functioning as Parts of Larger Units Only</i>	52
5.2.3 <i>A+N Phrases Functioning both as Complete Translation Units and as Part of Larger Translation Units</i>	55
5.3. A+N PHRASES WITH MORE THAN ONE TRANSLATION EQUIVALENT	57

5.3.1 A+N Phrases Whose Translation Equivalents are Synonymous	58
5.3.2 A+N Phrases Whose Translation Equivalents are not Synonymous	61
5.3.3 Two Special A+N Phrases as Parts of Translation Units	65
5.4 CONCLUSION	69
CHAPTER 6	
COMPARISONS WITH BILINGUAL DICTIONARIES.....	71
6.1 TWO DICTIONARIES USED IN THE COMPARISON	72
6.1.1 Introduction of A New English-Chinese Dictionary (Century Edition)	72
6.1.2 Introduction of English-Chinese Glossary of Legal Terms (Web version)	74
6.2 A COMPARISON OF THE CORPUS WITH THE NEW ENGLISH-CHINESE DICTIONARY (NECD).....	77
6.3 A COMPARISON WITH THE ENGLISH-CHINESE GLOSSARY OF LEGAL TERMS	85
6.4 CONCLUSION	90
CHAPTER 7	
CONCLUSION.....	91
7.1 RESEARCH PROBLEMS AND METHODS	91
7.2. MAIN RESULTS.....	93
7.3. RELEVANCE OF THIS STUDY AND FURTHER WORKS	95
REFERENCES	99
APPENDICES.....	105
APPENDIX 1: EMAIL FROM THE DEPARTMENT OF JUSTICE	105
APPENDIX 2: PERL 1	106
APPENDIX 3: PERL 2.....	107
APPENDIX 4: PERL 3.....	108
APPENDIX 5 : PERL 4	109
APPENDIX 6:10 SAMPLE EXTRACTION OF <i>CONCLUSIVE EVIDENCE</i> AND ITS CHINESE EQUIVALENTS .	110

Index of Tables, Examples, Figures and Pictures

Tables

TABLE 1: 30 A+N PHRASES	38
TABLE 2: 20 A+N PHRASES WITH ONE CHINESE EQUIVALENCE	48
TABLE 3: 13 A+N PHRASES WITH THE SAME TRANSLATION EQUIVALENTS ARE ALSO TRANSLATION UNITS	51
TABLE 4: A+N PHRASES AS PARTS OF LARGER TRANSLATION UNITS.....	53
TABLE 5: A+N PHRASES BOTH AS TRANSLATION UNIT AND AS PARTS OF TRANSLATION UNITS	55
TABLE 6: 5 A+N PHRASES WHOSE TRANSLATION EQUIVALENTS ARE SYNONYMOUS	58
TABLE 7: 3 A+N PHRASES WHOSE TRANSLATION EQUIVALENTS ARE NOT SYNONYMOUS:	61
TABLE 8: COLLOCATES OF <i>MEDICAL OFFICER</i>	64
TABLE 9: TRANSLATION EQUIVALENTS OF <i>GOOD ORDER</i>	67
TABLE 10: TRANSLATION EQUIVALENTS OF <i>RESIDENTIAL CARE</i> :	69
TABLE 11: THREE A+N PHRASES LISTED AS SUBENTRIES IN THE NECD.....	78
TABLE 12: THREE A+N PHRASES LISTED AS EXAMPLES IN THE DICTIONARY	79
TABLE 13: 24 A+N PHRASES NEITHER AS SUBENTRIES NOR AS EXAMPLES	80

Examples

EXAMPLE 1	44
EXAMPLE 2	44
EXAMPLE 3	46
EXAMPLE 4	47
EXAMPLE 5	51
EXAMPLE 6:	60
EXAMPLE 7	63

Figures

FIGURE 1: CONCORDANCE OF <i>GOOD ORDER</i> :	65
FIGURE 2: CONCORDANCE OF <i>RESIDENTIAL CARE</i>	67
FIGURE 3: DICTIONARY CONSULTATION OF TRANSLATION UNITS	84

Pictures

PICTURE 1 : THE WEB VERSION OF THE ENGLISH-CHINESE GLOSSARY OF LEGAL TERMS	75
--	----

Chapter 1

Introduction

Translation units are the basic units in translation. This thesis examines translation units from the perspective of corpus linguistics. It analyses translation units by considering their equivalents. The central theme of this dissertation is that translation units and equivalents can be extracted from parallel corpora and can be used in a bilingual dictionary or bilingual lexicon or translation database for the benefit of translators. The study on translation units reported in this dissertation was based primarily on the observation of 30 sample adjective + noun phrases extracted from an English-Chinese parallel corpus: the Hong Kong Legal Document Corpus (HKLDC) held at the Centre for Corpus Linguistics, the University of Birmingham. This first chapter of the dissertation presents the background of the study, specifies the problem of the study, describes its significance and gives the organisation of the dissertation.

According to Tognini-Bonelli (2001), there are two typical methodologies in corpus linguistic studies-- *corpus-based* and *corpus-driven*. The *corpus-based* approach refers to a method that “avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study (ibid: 63). The *corpus-driven* approach means a method “where the linguist uses a corpus beyond the selection of examples to support linguistic argument or to validate a theoretical statement”(ibid: 84). In a corpus-based approach, the corpus is seen “as a repository of

examples to back pre-existing theories or a probabilistic extension to an already well defined system” (ibid, 84). The corpus data are adjusted to quantify a theory. However, in a corpus-driven study, the corpus data will be used as a whole to provide the evidence for analyses or theoretical statements that may not be covered by existing theories. This dissertation has adopted the corpus-driven method. It aims to use an English-Chinese parallel corpus to uncover the new characteristics of translation units and posit and test hypotheses. All the sample data will be counted instead of being ignored.

1.1 Background

Globalisation demands quick, accurate translation. However, traditional dictionaries have shown their deficiencies in meeting this requirement. The traditional bilingual dictionaries have inherent deficiencies, especially when it comes to helping those translating from their native language into a target language they know less well (Teubert, 1999, 2002). For a long time, bilingual lexicographers have been looking for a dictionary which can provide “real lexical units of the target language which, when inserted into the context, produce a smooth translation” (Zgusta, 1984: 147). This research is undertaken in this context. It aims to apply corpus linguistics theory to improve bilingual lexicography and help those translators who need to translate from their native languages.

One of the main reasons for the inherent defects of traditional bilingual dictionaries is, according to corpus linguists, that traditional bilingual dictionaries take the single word in isolation as the standard entry and provide a number of decontextualised equivalents. The problem of using the single word as the standard lemma is that it will frequently result in

ambiguity (Sinclair, 2004:25). Since the most frequent words of a language are typically polysemous, for each word that they seek to explain, describe, and define, traditional bilingual dictionaries usually provide several equivalent options in the target language.

“The more polysemic a word in the source language is, the more translation equivalents it may have in the target language and the more there may be a need on the user’s part for such disambiguating information” (Swanepoel, 2003:69).

The dictionary user’s need is to know the exact meaning of each translation option of a word, and the precise grammatical, collocational, stylistic, discoursal and genre-specific conditions of its use. It requires the dictionary to provide very subtle differences in meaning of any translation options when a translator is translating from his/her native language. Traditional bilingual dictionaries, for various reasons, can seldom give proper instructions of where and when to use which option, and dictionary users therefore have no way to select the proper equivalent, especially when they have no intuition or large vocabulary in the target language. In a word, single words are ambiguous.

This thesis argues that parallel corpora provide a solution to these problems. Parallel corpora are collections of authentic texts and their translations into one or more target languages; they contain the practice of many experienced translators. Parallel corpora reflect what translators do. The main advantage of using parallel corpora is that they can provide contextual meaning for each word. There are many translation units and translation equivalents which can help the disambiguation. They help their users or translators by providing needed discourse, stylistic, and grammatical information. Parallel corpora help to discover what the context is, if we do not know the correct translation equivalent.

The essence of corpus linguistics is to look at words in context. Corpus linguists argue that words will be disambiguated if they are looked at together with their context. Therefore, larger language units rather than isolated single words should be studied. The practice of professional translators also shows that the texts are not translated word by word, but by units often larger than a single word. Teubert (1996, 2001, 2002) proposes that there is a smallest unit in translation, larger than single words, and calls it the ‘translation unit’ (details are explained in Chapter 2). He argues that translation units and their translation equivalents as found in parallel corpora can “complement traditional translation aids, such as printed dictionaries, term banks and even translation memories” (Teubert, 2002:211). However, his study was based on manual exploration of comparatively small corpora. He stated that the evidence from his comparison was limited and a larger parallel corpus would be needed to find enough occurrences of each translation units.

In order to verify his claim, Teubert led the *TranslationBase* project which aimed to extract translation units and their equivalents in parallel corpora. Building an English-Chinese parallel corpus, Chang et al. (2005) used statistical approaches aiming to extract translation units (multi-word units) based on Teubert’s definition. The idea was only tested on a small sample of these documents (500 aligned sentence pairs); this process yielded a candidate list which required a large amount of human evaluation (Chang et al, 2005: 139). However, without understanding which is the source language (Chinese or English), the majority of what Chang et al (2005) claimed as multi-word units are in fact single English words matching Chinese words consisting of more than one characters. They were not the expected larger translation units. Chang’s work later was continued by two Chinese scholars, Sun Le and Qu Weiming, who aligned the whole corpus (on which the present thesis is based). They

then used statistical methods to produce thousands of multi-word unit candidates according to different syntactic categories (Noun+Noun, Adjective+Noun and other structures). These candidates were later validated by Lianzhen He from ZheJiang University, China. She claimed the result was rather disappointing because only about three hundred of them could be regarded as multi-word units (Personal communication with He Lianzhen, 2003).

This overall unsuccessful attempt at extraction by pure statistical methods suggests that we need to describe the linguistic features before we can automatically and accurately extract them. In other words, the characteristics of translation units must be studied first before we try to make the computer extract them automatically. This dissertation seeks to show what the translation units and their equivalents contained in the parallel corpus should look like, and describes their characteristics.

1.2 Aim

Although Teubert has not shown what these translation units should look like, and Chang et al. could not automatically extract them, this dissertation illustrates that this concept of the unambiguous translation unit is very important because it describes translation equivalence in such a way that the problem of ambiguity disappears. The translation equivalents of these translation units document the practice of the community of translators of a given language pair. As this dissertation will show, they are more comprehensive and accurate than could possibly be found in a traditional bilingual dictionary. In addition, they provide a solution to the problem of ambiguity in translation. If this larger unit replaces the single word in translation, it may change the profile of our present bilingual dictionaries. Once extracted,

these ready-to-use units and their equivalents can be used and reused in further translation, and they will be especially helpful to the translators who need to translate from their native language. In addition, they will significantly increase translation speed and quality.

The first objective of this research is to clarify what the translation unit and the translation equivalent are. This dissertation will use the definition given by Teubert who defines them from the corpus linguistic point of view. This makes it different from other terms which are also called *translation units* or *translation equivalents* but which have not been strictly defined or have been defined from different points of view, such as on the basis of statistical probability by computational linguists (Wu, 1995; Yamamoto et al., 2001; Ribeiro et al., 2001; Aramaki et al., 2001). Based on Teubert's definition, a hypothesis is formulated that a translation unit has, in principle, only one translation equivalent in the target language. This hypothesis will be tested and verified through the discussion of translation units and their equivalents.

The second research objective is to demonstrate how translation units and translation equivalents can be extracted from parallel corpora. Inspired by pattern grammar (Hunston and Francis, 2000), this study has tried to extract translation units based on the syntactic pattern Adjective + Noun (A+N). This dissertation will discuss how the selection process of translation units and translation equivalents is carried out. It can only show some samples of this kind of translation unit and their equivalents based on this pattern of extraction to analyse their characteristics and properties. It is beyond the scope of this dissertation to discuss the automatic extraction of all the translation units.

The third objective is to show that the equivalents of these translation units yielded from parallel corpora are better than those provided by traditional dictionaries. In order to verify that the translation equivalents yielded from the parallel corpus will be preferable to those translation equivalents found in traditional dictionaries, a comparison will be made between corpus evidence and dictionary evidence. There are two dictionaries used in the comparison: one is a general dictionary; the other is a specialized legal glossary. The findings will indicate that the contextual translation units and equivalents embedded in the real text are hardly ever found in our present bilingual dictionaries. What is needed is something to supplement these bilingual dictionaries.

In summary, the aim is not to realize the automatic extraction of meaningful translation units, but to extract some sample translation units and their translation equivalents and to provide evidence that they are better than those yielded from bilingual dictionaries.

1.3 Organization

This thesis is organized in the following way: Chapter 2 provides the theoretical background of the concept of the translation unit in corpus linguistics. This originates from the concept of an extended unit of meaning in monolingual context. The extended unit of meaning and translation units are the same notions but from different perspectives.

Apart from explaining what the translation unit is in the sense of corpus linguistics, this thesis will show how the translation units and translation equivalents can be extracted. Chapter 3

introduces the 10-million-word Hong Kong Legal Document Parallel Corpus (HKLDC), on which my research is based. Chapter 4 describes how the sample translation units and their equivalents have been extracted, including the algorithm involved.

Both Chapter 5 and 6 are the main body of this thesis. Chapter 5 focuses on analysing the translation units yielded from analysing the 30 A+N phrases and their translation equivalents. In Chapter 6, these equivalents are compared to those provided by two traditional English-Chinese dictionaries, one of which is generic, while the other is a specialised dictionary. If the translated texts are the aim that translators are targeting during rendition, the more the same translation equivalents found in the dictionaries, the better.

Finally, Chapter 7 concludes the whole dissertation by upholding the value of parallel corpora as a powerful reference tool in translation. It also points out the further work needed to be done.

Chapter 2

Translation Unit: From Theory to Practice

This chapter is organised as follows: since the concept of the translation unit used in this research derives from the concept of a unit of meaning in monolingual context, Section 2.1 starts by introducing the corpus linguistic view of the unit of meaning. It focuses on the contributions made by John R. Firth (1890-1960) and John M. Sinclair (1933-) to the contextual theory of meaning which argues that the unit of meaning is above the individual word. Section 2.2 defines the notion of the translation unit and the translation equivalent used in this dissertation. The relationship between the translation unit and the unit of meaning is also considered in this section. Section 2.3 discusses several earlier studies. The difference between translation unit studies and corpus-based translation studies using parallel corpora is also discussed in this section.

2.1 The Unit of Meaning in Corpus Linguistics

The unit of meaning is regarded as “the starting point of the description of meaning in language” (Sinclair, 2004: 24). From the monolingual perspective of the source language, the translation unit is in fact a kind of unit of meaning. Its equivalence in the target language forms another unit of meaning. However, what constitutes the unit of meaning has long been disputed. Traditional linguists and lexicographers have tended to regard the individual word as the unit of meaning. Corpus linguists, however, propose that the meaning of a lexical word can only be understood through the contexts in which it occurs; therefore, the unit of meaning

should be above the individual word. This section will introduce the development of the corpus linguistic theory of contextual meaning and extended units of meaning from which the concept of the translation unit in this thesis is derived.

2.1.1 Word – Traditional unit of meaning

Usually, people read or listen to texts to understand their meaning. Meaning is thus the core feature of natural language. However, meaning has often been seen as secondary or even irrelevant to linguistic study. As Tognini-Bonelli (2001) has pointed out, traditional linguistics has had little relevance to the study of meaning. Where meaning is an issue, the word, the continuous string of letters separated by spaces or punctuation marks, always remains the basic unit of meaning of a semantic system and the study of meaning is mainly confined to the single word.

Lexicography has reinforced the concept that the word is the basic unit of meaning. Traditional dictionaries alphabetically list the individual words as lemmas and describe the range of meanings of a single word, thus confirming the equation “word = unit of meaning” (Sinclair, 2004: 25). Although some lexicographers (Geeraerts, 2003; Burkhanov, 2003) have commented that words do not exist in isolation and have tried to include some larger entities such as compound words and idioms in dictionaries, their normal practice is still at the level of describing the individual meaning of a given word. The meanings in the traditional dictionaries are still meanings ascribed to single words in isolation.

The unit of meaning is a semantic unit. In their discussions of lexicology, Lyons (1968, 1977), Cruse (1986) and Aitchison (1987) paid attention to the study of meaning (semantics) but not to the syntagmatic importance of context. They focused on paradigmatic relations between words (for example, antonymy or synonymy) but not on the relationship of co-occurrence. None of them worked with empirical data (i.e. a corpus) when they discussed the concept of the semantic unit. In their view, the single word was the semantic unit.

More recently, some lexicographers (such as Moon, 1998) have noticed that fixed expressions can only be fully understood if they are considered in the context of the texts in which they occur, and has suggested that new, use-centred models are required. This suggests that the corpus linguistic view of contextual meaning and extended unit of meaning has come to be more widely accepted. However, the real challenge to the traditional view of word as the unit of meaning can be traced back to Firth. Section 2.1.2 will explain Firth's view of the unit of meaning.

2.1.2 The Foundation of Changing the View—Firth's Theory about Meaning

Corpus linguistics replaces our traditional notion of the word as the core semantic unit by the notion of the extended unit of meaning. Corpus linguists argue that a unit of meaning may, in some cases, be a single word, but in more cases, it will be more complex. It is J. R. Firth (1890-1960) who "laid the theoretical foundation of a contextual theory of meaning which is central to our present-day view of corpus work" (Tognini-Bonelli, 2001: 157).

Firth argues that the major task of descriptive linguistics is to "make statements about meaning" and that meaning can be stated in terms of natural language (Firth, 1957b: 190-192). His contextual theory of meaning emphasizes the importance of context: the context of surrounding words. According to this theory, the meaning is carried not by the word itself, but by the word embedded in its context. His famous dictum is that a word is characterized by the company it keeps. Because the meaning of a word lies in its use, and use does not exist in isolation; "... you shall know a word by the company it keeps" (Palmer, 1968: 179), not by consideration of the word in isolation. The meaning of a word should be interpreted together with its context, and a word in a new context is in fact a new word. "... the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously" (Firth, 1935:37 cited in Stubbs, 1993:9; Firth, 1957b:7). Firth's context means "a level of language description where the limitless complexity of the non-linguistic environment is organized into linguistically relevant categories" (Sinclair, 1991:171).

Firth put forward two revolutionary notions which have had considerable impact on modern linguistics: collocation and colligation. "Collocations of a given word are statements of the habitual or customary places of that word in collocational order...not in any grammatical order" (Firth, 1957a:181). According to Firth, meaning by collocation is the result of the fact that many words regularly (what we would now call *significantly*) occur with some words within a particular context. From today's point of view, Firth may not have had the intention of using the concept of collocation to define the larger phrasal units as units of meaning, but we can assume that his concept of collocation surely includes complex items.

Another revolutionary notion initiated by Firth is colligation. Colligation is “... the interrelation of grammatical categories in syntactical structure” (Palmer, 1968: 169) or, “...in a slightly wider sense, a pairing of lexis and grammar” (Stubbs, 2002:65). Meaning by collocation is placed at the lexical level, while meaning by colligation is placed at the grammatical level. For example, nouns are usually preceded by determiners and adjectives. “Colligations of grammatical categories related in a given structure do not necessarily follow word divisions or even sub-divisions of words” (Firth, 1957a: 182). What is more, the collocational level must be matched by the colligational level. Francis (1993: 141) has discussed the mixed characteristics of collocation and colligation in phraseological units.

In the days dominated by transformational research, Firth’s view did not attract the attention it deserved. Because computers and electronic corpora were not available, he was unable to fully explore “words in context” and meaning at the collocation and colligation levels. It is John M. Sinclair who has brought real changes in our views of natural language and our perceptions about unit of meaning. This will be explained in section 2.1.3.

2.1.3 Lexical Item — the Extended Unit of Meaning

Sinclair has developed much of the theoretical framework of what is now called corpus linguistics. He was the project leader of the famous Cobuild project which has promoted both linguistic applications as well as linguistic theory. For Sinclair, there are two conflicting principles in language organisation: *the open-choice principle* and *the idiom principle* (Sinclair, 1991, 1996, 2004). By *the open-choice principle*, he means that words can be put

together randomly, while by *the idiom principle* he means that words co-select each other and tend to co-occur. He argues that the idiom principle is the dominant way that text is formed.

Collocation, in Sinclair's theory, is the illustration of the idiom principle, and it is the co-selection of words: "Collocation is the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991:170). Stubbs (2002: 24) has further pointed out that the collocation is "a lexical relation" between two or more words which have a tendency to co-occur within a certain span. A collocation is largely a node-collocate pair. The node word is the word which is under examination, and a collocate is a word that significantly co-occurs with the node. Node and collocate are exchangeable according to the different study purposes. Collocation is not syntactically directional because a collocate can be either on the left or right of a node. However, collocation is directional in the sense of frequency — *downward collocation* if the node is more frequent than collocate, and otherwise *upward collocation*.

From the surface to the core, there are four layers of relationship of language co-occurrence: *collocation*, *colligation*, *semantic preference* and *semantic prosody* (also called *discourse prosody* by Stubbs (2002)). Both *collocation* and *colligation* are concrete, indicating the relation between individual words while *semantic preference* and *semantic prosody* are rather abstract, indicating the semantic environment. *Semantic preference* is the relation between a lemma and a set of semantically related words while *semantic prosody*, extended over more than one unit in a sentence, can express the speaker's attitude.

Based on the idiom principle and four levels of language co-occurrence, Sinclair (1996, 2004) argues for “extended units of meaning”¹. He suggests that units of meaning are “largely phrasal”, and that the idea of a word carrying meaning on its own needs to be discarded because only a few words (for example “in the enumeration of flora and fauna”(Sinclair, 2004:30)) are selected independently. He calls these extended units of meaning “lexical items” (Sinclair, 1996, 2004). Compounds, phrasal verbs, idioms, fixed phrases, variable phrases, clichés, proverbs, and technical terms and jargons as well as collocations are all units of meaning because in these patterns the whole expression has a meaning which is different from the sum of their component individual meanings. Even in a collocation where each individual word still seems to mean what it normally means, "there is usually at least a slight effect on the meaning"(Sinclair, 1996:80).

Sinclair has thus replaced the individual word in the traditional view of the standard unit of language with a new concept of the extended unit of meaning. Although it is not certain whether the lexical item is the only form of a unit of meaning or not, it is certain that a lexical item is the main form of the extended unit of meaning. Sinclair (1991, 2004) assumes that very large corpora are needed to provide evidence for larger units of meaning. Concordances (KWIC) are the standard modes of data presentation which can show these units of meaning. In many of his works (e.g. Sinclair, 1996, 1998 and 2003), Sinclair discusses the theory and methodology of identifying units of meaning based on corpus data. In practice, the success of the COBUILD project has validated his theory.

¹ After Sinclair, Michael Stubbs (2002) has also illustrated in detail why individual words are not always the unit of meaning.

2.2 The Translation Unit in Corpus Linguistics

2.2.1 Translation Unit and Translation Equivalent

The section above has presented the corpus linguists' view about meaning and has mentioned the new concept of the extended unit of meaning as proposed by Sinclair. However, the unit of meaning is a concept for monolingual settings. In the bilingual context, which is the context of my study, I replace the unit of meaning by what Wolfgang Teubert calls the translation unit.

It is widely known that professional translators do not translate texts word by word. They are normally translating large chunks, for example a collocation or a unit of meaning as a whole. Teubert (1996, 2001, 2002) calls these chunks “translation units”. In his view, translation units are centred on lexical words. “Lexical unit and relevant context together form the translational unit” (Teubert, 1996: 256). What is more, these units are ideally monosemous. Later, Teubert (2001, 2002) developed his theory of translation units and defined a translation unit as “the smallest monosemous unit in translation”. It comprises all those words in context which should be translated as a whole.

The translation unit is a contextual semantic unit. It is not judged by syntactic structure (for example, a sentence or clause) but by its meaning. It can be a continuous string of words or words not adjacent to each other. It can be a phrase, a clause, or a sentence. In this thesis, I will not discuss cases where the translation unit extends beyond the sentence boundary. For one reason, it is difficult to extract a translation unit and its translation equivalent if they are above sentence level due to the limitations of sentence alignment. For another reason, the aim

of the study of translation units and of translation equivalents is to reuse them for future translations, but in reality the probability of the repetition of a unit extending beyond a simple sentence is not high. The translation units discussed in this thesis are mainly at the phrasal level.

“Translation units, consisting of a single word or of several words, are the minimal units of translation” (Teubert, 2001:144). Teubert has not yet excluded single words from translation units but he claims that most translation units are multi-word units which are larger than single words. Single words are usually polysemous and ambiguous from the target language perspective. If a word is always translated as a certain fixed equivalent in the target language, it can of course be a translation unit, but a glance in any bilingual dictionary will confirm the fact that most single words have more than one equivalent. This phenomenon indicates that the majority of translation units are larger than a single word. Sometimes, even if a word seems to be unambiguous in a monolingual context, more equivalents for it may be found in another language. A famous example given by Teubert (2003) is the English word *bone*. It is consistently described as monosemous in monolingual dictionaries but it has two translation equivalents in German according to whether *animal bone* or *fish bone* is being referred to.

Teubert (2001) calls the equivalence of a translation unit in the target language a “translation equivalent”. He advocates that the translation equivalent is a paraphrase of a translation unit, namely a paraphrase in the target language. “The meaning of a translation unit is its paraphrase, that is, the translation equivalent in the target language” (Teubert, 2001:145). In other words, a translation equivalent is the meaning of a translation unit, but in the target

language. A translation unit is, by definition, monosemous, which means that it will have only one translation equivalent in the target language. If it has more than one translation equivalent, these translation equivalents should be synonymous and should be able to replace each other. If there is more than one target language equivalent, and they are not synonymous, then the source language expression is not yet a translation unit, and therefore it has to be extended until it becomes a translation unit. In other words, one or more context words have to be added to it until it becomes, from the target language perspective, unambiguous.

It is worth pointing out that the term *translation unit* and *translation equivalent* are similar concepts. A translation equivalent is the equivalent of a translation unit. Scholars in computational linguistics (Wu, 1995; Yamamoto et al., 2001; Ribeiro et al., 2001; Aramaki et al., 2001) and translation studies (Malmkjær, 1998) have used the concept of translation unit or translation equivalent as well, but they either have not strictly defined these two concepts or they used them as more or less statistically derived units, not as concepts linguistically defined in translation. Their translation units are not necessarily the semantic units. Baker (1992) mentioned translation equivalence at different levels (from word level to textual level and pragmatic level). Similar to Teubert, she proposes that it is the unit of meaning which is more often larger than the single word that is translated. She claims that the translation is influenced “by a variety of linguistic and cultural factors” (ibid: p6), however, she has not further investigated whether these factors should be included in the translation unit or not.

2.2.2 The Relationship of Translation Unit and Unit of Meaning

The translation unit has been defined as the bilingual counterpart of the unit of meaning. The translation unit is defined from the perspective of the target language. The unit of meaning and the translation unit are similar to each other in that both of them are, ideally, unambiguous semantic units. Their difference lies in the fact that the unit of meaning is based on the monolingual perspective while the translation unit is based on the perspective of another language.

In other words, what is a translation unit depends on the target language. An expression that may have only one meaning in the source language may have two or more translation equivalents in the target language. For example, the phrase *medical officer* is unambiguous in English from the perspective of the Hong Kong Legal Document Corpus used in this thesis. It refers to a government doctor, and a native speaker of English is unlikely to misunderstand it. But in the Chinese text, *medical officer* is polysemous. It can be translated as 公职医生 /*gong zhi yi sheng* and 政府医生 /*zheng fu yi sheng*. These two translation equivalents are not synonymous; therefore, more context words are needed to form two monosemous translation units. The translation units therefore have to be extended. How this is done will be analysed in Chapter 5. What is more, “What is a translation unit in relation to one target language does not have to be one in relation to another” (Teubert, 2002: 212). If the target language is different (e.g. it is not translated into Chinese but into Japanese), we may find different translation units accordingly. If the phrase *medical officer* is translated into a language where we find only one translation equivalent for it, then it would count as a translation unit.

In the monolingual context, a unit of meaning is the node plus all of its collocates. Similarly, a translation unit is composed of the node under investigation plus its collocates which are relevant for disambiguation. In a monolingual context, any word can be a node and thus form a unit of meaning, but in the bilingual context, the node of the translation unit should usually be a lexical unit because lexical words are usually translated. Some grammatical words, such as *the* in English, have no counterpart in other languages, including Chinese. These words can help translators to understand the original texts, but they cannot be the core part of a translation unit. Statistically, if every item in a translation unit has a score as its value, the score of the grammatical items should not occupy a high percentage in the whole.

Translation units may inherit the characteristics of the unit of meaning (lexical items) described by Sinclair (1991:111): indeterminate extent, lexical variation, syntactic variation and variation in order.

2.2.3 Criteria of Identifying Translation Unit in this Paper

Since the translation unit is defined from the target language perspective, it may be a good starting point to identify a translation unit from the translation equivalent in practice. Theoretically, two criteria have been used to find a translation unit: semantic relevance and the frequency of recurrence.

One criterion is that there should be semantic relevance among the lexical units in the translation unit. In other words, all the words of a translation unit should form one expression. If there are two things expressed in a unit, the unit will not be the smallest. For example,

residential care in *residential care home* is a modifier of the word *home*, even though in other contexts it can occur as an expression on its own. The whole phrase *residential care home* is semantically one thing and has been translated as a whole as 安老院/*an lao yuan*. The example will be further discussed in Chapter 5.

From the target language point of view, a translation unit will always be translated into only one translation equivalent. In any sizable corpus, if a translation unit is often repeated in the source text, its translation equivalent should also occur in comparable frequency. A translation unit can be identified from the recurrence of its translation equivalent.

The problem is how to keep the unit as small as possible. Is *Hong Kong Special Administrative Region* (translated in Chinese as a whole: 香港特别行政区/*xiang gang te bie xing zheng qu*) one translation unit or should *Hong Kong* and *Special Administrative Region* be two independent translation units?

The criterion used in this paper is to judge whether this unit, although it may include several lexical units, can or cannot be further reduced to smaller unambiguous units. If, when we reduce the unit to several smaller lexical items, at least one lexical item will be polysemous, then the unit cannot be broken down. But if all of the smaller units are independently monosemous, the whole translation unit will not be the smallest monosemous unit. For example, *special administrative region* is monosemously translated as 特别行政区/*te bie xing zheng qu*, and it can be regarded as a translation unit; *Hong Kong* is always monosemous in Chinese as well. Therefore, they are two different translation units. But *the court of final*

appeal is different. *Court* is polysemous and can be either translated as 法院/*fa yuan* in the *Court of Final Appeal* or as 法庭/*fa ting* in the *Court of First Instance*. Although *final appeal* (translation as: 终审/*zhong shen*) can be regarded as a translation unit, *the court of final appeal* can only be regarded as one translation unit since the other unit *court* is not monosemous. The practice of keeping translation units as the smallest units suits Occam's Razor (also called the law of economy) which means the simplest is the best.

2.3. Translation Unit and Parallel Corpora

2.3.1 Types of Parallel Corpora

Corpora are collections (usually electronic ones today) of texts. "A 'parallel corpus' is a bilingual or multilingual corpus that contains one set of texts in two or more languages" (Teubert, 1996: 245). There are altogether three types of parallel corpora, and their functions are different according to their different construction.

The first type is the normal parallel corpus. This type contains only texts of language A (usually source language) and their translation into language B (target language). The HKLDC corpus belongs to this type.

The second type is the reciprocal parallel corpus. It contains not only the source texts in language A and their translation in language B, but also source texts in language B and their translation in language A.

The third type contains only translations in different target languages. This type may be bilingual or multilingual; for example, the corpus containing French and German translations of Plato's *Republic*. It may also be monolingual; for example, a corpus contains only seven English versions of Plato's *Republic*.

2.3.2 The Translation Unit Study vs the Corpus-based Translation

Studies

Since this dissertation uses a parallel corpus to analyse the basic unit in translation, it is better to clarify the difference between this study (called *the translation unit study*) and the translation studies based on corpora. Translation studies have more and more been regarded as a scientific discipline and scholars in translation studies are more frequently conducting their researches by using corpora. This is known as *corpus-based translation studies*. Corpus-based translation studies are not only based on parallel corpora (Ebeling, 1998), but also on monolingual (Bowker, 1998) or comparable corpora (Baker, 1995, 1996; Laviosa-Braithwaite, 1996). If parallel corpora are used, these translation studies tend to work with the third type of parallel corpora which contain only the translation texts.

Both the studies conducted in this dissertation and corpus-based studies work with translated texts, although this study only worked with the first type of parallel corpora where both source language and target language are contained. Scholars in translation studies use translation corpora to explore an ideal translation which can minimise the inevitable distortion of the original text in the sense of information (message), spirit and elegance of the original

language. They are interested in whether two expressions are equivalent in meaning or not, and to what extent.

This study does not consider how the texts have been translated. The translated texts are regarded as already good enough input for the analysis according to the translation criteria. This is based on the simple fact that translation errors are less often repeated than successful translations. Therefore, frequency will filter out any translation errors. These studies aim to find how the original meaning has been represented in the target language unit by unit (translation units). The task is to identify these monosemous translation units in the original language, to extract their equivalents in the target language, and to reuse them either by designing a new bilingual dictionary or creating a translation database.

Corpus-based translation studies are interested in how equivalence might be achieved and what kind of equivalence can be achieved, and in what context; translation unit studies are interested in the alignment of translation units and their equivalents in a given parallel corpora, and how these equivalents can be re-used by other translators in the future translations, especially by those translators who have to translate into a non-native language where their intuition is often insufficient.

2.4. Summary

This chapter has discussed the corpus linguistics view of meaning both in the monolingual context (unit of meaning) and in the bilingual context (translation unit). Translation units as the smallest monosemous unit in translation are very useful for bilingual lexicography and

Machine Translation. Teubert (2001) maintains that parallel corpora are repositories of the translation units and their equivalents. The following chapter will describe the parallel corpus – the Hong Kong Legal Document Corpus (HKLDC) – from which the translation units and their translation equivalents were extracted.

Chapter 3

The Hong Kong Legal Document Corpus (HKLDC)

Translation units and their equivalence in this thesis were extracted from the Hong Kong Legal Document Corpus (HKLDC). The HKLDC has been compiled by the former Centre for Corpus Linguistics at the University of Birmingham. It contains approximately 10 million words Hong Kong bilingual laws. Compared to the large monolingual corpora such as BNC (British National Corpus, 100 million words) and BoE (Bank of English, 450 million words and it is still growing), the size of the corpus is rather small, but it provides translations of high quality and enough recurrence of translation units.

This parallel corpus was compiled as the basis for a Chinese-English *TranslationBase* project which was started in 2000, led by Prof. Teubert at the University of Birmingham, with partial financial support from HarperCollins up to 2003. Three Chinese visiting scholars (Dr. Baobao Chang, Dr. Le Sun, and Dr. Weimin Qu) were involved in compiling, pre-processing and aligning the corpus. Before the research for this thesis, the corpus had been aligned and the alignment had been roughly checked. The work carried out by previous scholars saved much time which would otherwise have been spent on corpus compilation. However, due to the discontinuity of the work through the conditions under which it was compiled, some information of the corpus was lost and it is impossible to trace it altogether accurately. For example, it is unclear from which websites the texts were downloaded originally. This chapter will try to introduce the corpus as fully and accurately as possible.

3.1 Texts

The HKLDC contains the statutory laws issued by the Department of Justice of the Hong Kong S.A.R. Government before 2001 (inclusive). Most of them were issued on 30 June and 1 July 1997 when Hong Kong sovereignty was handed over from Britain to China. The corpus does not contain any laws issued after 2001. Its size is more than 10 million words (approximately 5.6 million English words and 4.6 million Chinese characters).

The Chinese text has fewer words than English. One of the reasons is that the Chinese language has fewer determiners and grammatical words than English. Chinese words are invariant and therefore a lot of space is saved. For instance, there is no article before a noun in Chinese; Chinese does not have obligatory tense markers in the same way English does, where sometimes several words are needed to indicate them.

It appears that all the texts in the HKLDC were downloaded from the Internet in 2001. Today, all the documents in the corpus can be found in the online information system, known as the Bilingual Laws Information System (BLIS) of the Department of Justice (<http://www.justice.gov.hk>). The Department of Justice is responsible to maintain and update this online database. Detailed information on BLIS can be found in Webster et al (2002) and Kit et al (2004). The difference between the texts in the HKLDC and in the BLIS is that the texts in the HKLDC are only part of the bilingual statutory laws of Hong Kong, and they were issued before or in 2001. The texts in the BLIS are the complete collection or archive of the statutory bilingual laws of Hong Kong, originally encoded in another format (the Lotus

Notes). They are often updated by the technical unit of the Hong Kong S.A.R. Department of Justice.

The Chinese text of Hong Kong bilingual laws was originally written in traditional Chinese characters. It was only at the late stage of alignment that these traditional Chinese characters were automatically converted into simplified characters because most Chinese scholars in this project were more familiar with simplified Chinese and because the software used for segmentation and tagging works better with simplified characters. In this conversion, something may have been lost. This is suggested in the information provided by the BLIS where a simplified Chinese version is provided but the government has not endorsed the simplified Chinese version.

The English text and Chinese text in the HKLDC have the same authoritative status. Like multilingual EU documents, there is no official declaration stating which text is the source language, and which the target language. Both of the English and Chinese texts are called “authentic texts”. In the case of conflict during implementation, it is the judge and the court which has the right to clarify which text is accurate. However, every effort has been made to identify the source language. For one thing, it is better to know which is source language (English or Chinese) because translation is regarded as a unidirectional activity, that is, from source language to target language. For another, the identification of the source language and target language could help extraction and analysis of translation units (e.g. to decide the starting point of extracting translation units especially when it is not known whether they are reversible or not).

Close analysis of the texts would indicate that English text is the source text. The English text seems to be more precisely stated than the Chinese one. Some information which the English text has painstakingly repeated has been omitted in the Chinese text although some of it can be deduced from the context. Also, when compared with a specialised Chinese-English corpus where Chinese is the source language, the HKLDC has no described typical linguistic characteristics for rendering Chinese into English. For example, the Chinese adverb 然后/*ran hou* is always omitted in the English version of Chan's (2002:3-22) TransRecipe (Cookbooks corpus), while in the HKLDC, 然后/*ran hou* almost always has an English counterpart: *thereafter* or *then*.

The strong evidence for this conclusion can be found from the history of the bilingual legislation in Hong Kong. Hong Kong statute law had been enacted only in the English language until 1987. After 1987, a Bilingual Laws Programme was launched to translate all laws which had been enacted in English. (For details of this programme please see Yen, 2001). This work of translation was completed just before 1 July 1997 when the sovereignty of Hong Kong was handed over back to the Chinese government. The following article provides the details:

"Up until the late 1980s, all the legislation in Hong Kong was enacted in English only." However after the signing of the Joint Declaration in 1984 it was obvious that Chinese would become the main language of Government in Hong Kong after the resumption of the exercise of sovereignty in 1997. In August 1986 the Royal Instructions were amended to enable laws to be enacted in Chinese, and in March 1987 the Official Languages Ordinance (Cap. 5) was amended to require all new legislation to be enacted in English and Chinese. This was in accordance with Government policy of providing a bilingual legal system for Hong Kong. The 1987 amendment to the Official Languages Ordinance (Cap. 5) also provided a mechanism for publishing authentic texts in Chinese, of Ordinances enacted in English only. As a result of this amendment a programme was launched (the Bilingual Laws Programme) to produce Chinese texts of all former laws that had been enacted in English only. That programme was completed shortly before 1 July 1997. Consequently all our legislation is now available in both English and Chinese. Both the English and the Chinese texts are authentic and are presumed to have the same meaning. Where a comparison of the texts discloses a difference of meaning that cannot be resolved by the rules of statutory interpretation ordinarily applicable, the meaning which best reconciles the texts, having regard to the object and purposes of the Ordinance, is required to be adopted....."

(from www.justice.gov.hk)

This conclusion has been confirmed by an email from the Department of Justice which drafts and issues the laws (see Appendix 1). It indicates that even if after the handover, the English text was normally drafted first, and then rendered into Chinese. “This is particularly true when the English text is drafted by Anglophone counsel who cannot write Chinese” (see Appendix 1). Bilingual draftsmen also prefer to do so as they have studied law in English, which is the working tool of the common law. Therefore, we can confidently maintain that English was the source language of the Hong Kong bilingual laws.

3.2 Annotation

The statutory laws of Hong Kong are divided into three categories: public ordinances (i.e. laws which concern the general public), private ordinances (i.e. laws which concern individual bodies, whether statutory or otherwise), and miscellaneous ordinances (i.e. laws which do not belong to either of the preceding categories). These ordinances have a very rigid numbering system in their divisions. These are, from greatest to least: Chapters, parts, sections, subsections, paragraphs and subparagraphs. Some divisions are identified by Arabic numerals, and some are marked by Roman numerals. In this corpus, this numbering system has been discarded and all the texts are put together into two files: English text *en_with_id.txt* and Chinese text *ch_with_id.txt*.

These two files are stored in two different forms: one is in plain texts only, the other is part-of-speech tagged. The English text in the HKLDC is tagged by TreeTagger (Schmied, 1994),

a language independent part-of-speech tagger developed at the Institute for Computational Linguistics of the University of Stuttgart. The Chinese text was segmented and part-of speech tagged by the Institute of Computational Linguistics, Peking University (www.pku.edu.cn). The segmentation of the whole Chinese text is done because there is no space between Chinese characters. It is difficult to use software if there is no space between them.

Texts are aligned by the “Vanilla” aligner (Danielsson and Ridings, 1997). The “Vanilla” aligner is a sentence-level alignment tool based on the algorithm proposed by Gale and Church (1993). It is worth pointing out that an aligned sentence in this corpus is not necessarily a linguistic sentence. It may be a sentence fragment or a clause, according to the convenience of the alignment. As a result, the sentence length is irregular; some are very long while some are very short. There are altogether 194,181 aligned sentences in HKLDC. The aligned sentences were manually checked and the accuracy of alignment is claimed up to 98% (Sun, 2003).

3.3 Advantages and Shortcomings

Hong Kong bilingual laws have several advantages for the study of translation units. The most obvious advantage is that we do not have to worry much about the problem of translation mistakes. The Bilingual Laws Programme had gathered the most famous experts in nearly all related fields from both Hong Kong and Mainland. They ensured the accuracy and consistency of the translation. Their translation represents the highest standard of the legal document translation.

Another advantage is that they are specialised parallel documents which are full of standardised bilingual terminology. The collocations extracted from the corpus are, to a large extent, terminological items. Therefore, there is a highly consistent relationship between translation unit and translation equivalents. This makes it easier to analyse the corpus.

Last, but not the least, these bilingual legal documents are comparatively easier for the alignment. The Chinese version of an ordinance follows the exact numbering system of its English counterpart. Accordingly, the English and Chinese texts are perfectly aligned with each other “in terms of Chapters (zhang 章), parts (bu 部), sections (tiao 条), subsections (kuan 款), paragraphs (duan 段) and subparagraphs (jie 节)” (Webster et al., 2002: 82). This feature also helps in manually checking the accuracy of the alignment result.

However, since the HKLDC is a specialised corpus containing only Hong Kong bilingual laws, it cannot reflect the general profile of Chinese and English. The representativeness is limited to a legal document domain within a limited number of users. The sampling of the English and Chinese in the HKLDC is restricted to the Hong Kong regional variety of English and Chinese.

3.4 Summary

In this chapter, a short introduction of the HKLDC is given from three aspects: the text it contains, the annotation and the characteristics. This corpus is comparatively smaller than many monolingual corpora, but it is consistent and of good quality in translation. The writer

feels confident that the source language and the target language have been identified and this is important because it can help to make the decision of where to start the extraction i.e. from English to Chinese. The details of the extraction will be discussed in Chapter 4.

Chapter 4

Methodology of Extraction Translation Units and Their Equivalents

Although today there are various software tools available for data extraction, there is not yet an ideal tool that would extract the desired English translation units according to the definition in this thesis and identify their translation equivalents in Chinese. Even ParaConc², a bilingual/multilingual concordance program for contrastive corpus-based language research, cannot help in yielding the exact translation unit and translation equivalent pair as defined in this dissertation. This indicates that pure statistical methods which these tools are based on will not work in this case therefore some linguistic factors must be considered in extraction. This chapter presents a methodology of using a syntactic pattern to extract translation units. The approach chosen is to choose Adjective+Noun (A+N) combinations as the start of extraction and manually identify their translation equivalents. The process is so far semi-automatic. Some specifically written Perl programs were used, but all the results extracted by automatic procedures need to be validated. Since the extraction results need human validation, they are called translation units and translation equivalents candidates. The validation process will be described in subsequent chapters.

This chapter explains why the A+N pattern was chosen, and how the translation unit candidates and their translation equivalents are extracted step by step. It will also propose an

² ParaConc is designed by Dr. Michael Barlow. It has been updated for several versions. The latest version is actually a bilingual/multilingual aligner as well as concordancer.

improved procedure for extracting translation units and their equivalent candidates which would reduce the task of intellectual validation.

4.1 Translation units and A+N pattern

The extraction starts from the English text. The translation unit, as discussed in Chapter 2, is in fact the unit of meaning in the source language. The source language is proved to be English (Chapter 3), therefore to extract translation units in the HKLDC is in fact to extract units of meaning in English text. Since corpus linguists believe that the majority of units of meaning are bigger than single words, what we mainly need to extract will be multi-word units, such as phrases.

Because computers so far can not understand meaning, the translation unit, as a semantic unit, cannot be automatically extracted yet. However, the computer can recognize syntactically defined patterns once the texts are part-of-speech tagged. As research in the field of collocation has shown, certain syntactic patterns are prone to be interpreted as units of meaning or, as in the case of translation, as translation units. A syntactic pattern is decided as a starting point for extracting translation units.

Among various types of patterns in Pattern Grammar (Hunston and Francis: 2000), the nominal pattern “adj N” is chosen in this research and the results of the extraction are called “A+N phrases”. There are three reasons why the decision to extract this pattern was taken. The first reason is that this pattern is most likely to be a unit of meaning. Preliminary research conducted for this study showed that most A+N phrases are potential translation

units in this parallel corpus. The second reason is that nouns and adjectives are the most frequent parts of speech in a text, therefore the high frequency of the occurrences of the phrases can be ensured. Biber et al (1999: 231) have observed that “nominal elements make up between a half and four-fifths of the text”. “The text” here means the general texts such as news reports, fictions, academic prose and so on whereas this study looks at legal texts. Nevertheless, this conclusion gives us further evidence that nominal phrases may be the right starting point for extraction. Moreover, in special language texts, Wright (1997:13) has found that “nouns” and “adjectives” are the most dominant parts of speech in the terminological units. Accordingly, the extraction of expressions conforming to the A+N pattern may yield good translation unit candidates. The third reason is that this pattern is a lexical pattern which will usually be translated instead of being omitted. This is important for us because we know their translation equivalents will always be somewhere in the corpus. Of course, other patterns will also yield translation unit candidates. As this dissertation is an exploratory study of limited scope, the translation unit approach will only be demonstrated for the A+N pattern.

4.2 Implementation of extraction translation unit candidates and their translation equivalents

The following describes a step-by-step implementation of the extraction procedures. Since the corpus had already been aligned, the extraction was based on its alignment. The aligned ID number in the parallel corpus formed the link point between the English and Chinese sentences. This means that for each English phrase, one can find its Chinese translation equivalent in the corresponding Chinese sentence.

- Step 1. Extract all A+N phrases from the English part of the parallel corpus; count their occurrence frequency in the corpus, and form an A+N combination list.
- Step 2. Select 30 A+N English phrases manually in the list. The frequency of the selected phrases is approximately 100 occurrences.
- Step 3. Extract the context of the 30 English phrases. For each selected phrase, extract 30 sentences in which this phrase occurs.
- Step 4. Extract the sentence ID numbers of the 30 extracted English sentences.
- Step 5. From English sentence ID numbers, extract the corresponding Chinese sentence with the same ID number from the parallel corpus.
- Step 6. Manually identify and extract the translation equivalents of the 30 A+N phrases.

Step 1 was completed with the help of Dr. Pernilla Danielsson. She provided a Perl program that extracts all the A+N combinations from the corpus. The algorithm of this program can be described as follows:

- (1) Open the tagged English texts in the corpus;
- (2) Read a word from the English texts;
- (3) If the label of this word is JJ (which is the tagging of adjectives), then check whether the label of the next word is NN (Singular form of Nouns) or NNS (plural form of nouns);
- (4) If yes, then an A+N phrase is founded; save this phrase in a file (File 1);
- (5) Repeat steps (2) — (4), until the end of English texts has been reached.

This program yielded more than 9,000 phrases which occurred three times or more. They were saved in a file (File 1). Since not every A+N combination is a meaningful translation unit, 30 phrases were selected (see Table 1). The first column in Table 1 gives the frequency of each phrase in the whole corpus. These 30 phrases were chosen because they appeared to

be promising candidates for translation units, and because they occurred around 100 times (the highest frequency was 105 times and the lowest was 88)³, which means they were not the most frequent ones but sufficiently frequent to permit reliable conclusions.

Table 1: 30 A+N phrases

Frequency	A+N Phrase	Frequency	A+N Phrase
105	straight line	94	legal adviser
104	legal officer	93	registered dentist
101	residential care	93	postal packet
101	criminal offences	93	good order
100	annual allowance	92	special category
99	long term	92	registered scheme
98	human remains	92	provisional registration
98	conclusive evidence	92	judicial trustee
97	written permission	91	internal combustion
97	public bus	91	final Appeal
97	personal representatives	90	necessary modifications
97	first column	89	rateable value
96	notifiable workplace	88	restricted licence
96	listed company	88	reasonable ground
95	light bus	88	medical officer

Thirty English sentences were then selected for each of these thirty phrases. The sentences were regarded as the contexts of each phrase. This step can be realized by a Perl⁴ program (Perl 1). (For the full algorithm of this Perl program, see Appendix 2). This Perl program is designed to extract only thirty sample sentences for each A+N phrase. However, in case there was an alignment mistake which would affect identifying the translation equivalence, some further sentences were extracted to ensure the occurrence was at least thirty.

³ These frequency figures are calculated by the Perl program which is used to extract the A+N phrases. However, these figures can only be used to ascertain roughly how frequently the phrases appear in the whole corpus. Different concordancing software may not yield exactly the same figures due to the different design of the query (e.g. some software queries may not include capital letters). For example, both ParaConc and the Perl program yielded 105 occurrences of the phrase *straight line*. However, Concapp, a free concordancing program by the [Virtual Language Centre](#) of the Polytechnic University of Hong Kong, yielded 106 instances of this phrase. Still, the results should be and actually are approximately the same. This study will use only the frequency figures yielded by the Perl program unless there is a fundamental difference between the figures in this study and the figures according to other software.

⁴ Perl is a simple program language. It works very well in processing words and texts.

Procedure Step 4 aimed to extract the English sentence ID number of those sample sentences which had been yielded by Step 3. Another Perl program (Perl 2) was used to complete this task. The algorithm for Step 4 is given in Appendix 3. As mentioned previously, the sentence ID number connects the English and Chinese sentences. In the HKLDC, the sentence ID number is always located at the beginning of each sentence, therefore the program regards it as the first word of each sentence.

In Step 5, the corresponding Chinese sentences with the same sentence ID number as the English sentences were extracted. Given that the file contains the English sentence ID number, the step can be implemented by another Perl program (Perl 3). The algorithm of Perl 3 is described in Appendix 4.

Identifying the translation equivalents of these A+N phrases in Step 6 was completed manually by the writer. This step was done manually because there was no procedure for automatic lexical alignment which could carry out this task with a satisfactory level of accuracy. The multilingual concordancer ParaConc⁵ helped to accomplish the above steps, especially in extracting the sample sentences for the thirty phrases. However, when an attempt was made to use ParaConc to realize Step 6, it did not work as expected since it cannot recognize the Chinese equivalents if the Chinese texts have not been segmented. Even after the segmentation, it could not reach the accurate translation equivalents of a whole phrase.

⁵ ParaConc can deal with parallel search and yield certain kind of translation equivalents by using the “Hot Words” function. The ParaConc software used was an old version—the first BETA version issued in 1996.

4.3 Discussion on the complete extraction

Step 3 described above aimed to extract thirty sentences for each A+N phrase. This was to ensure every A+N phrase would have about thirty occurrences. These thirty occurrences of each phrase provided useful information for the analysis of translation units. Sinclair (1991, 1996, 2003) suggests that when there are hundreds and thousands of concordance lines yielded, we should study them from the beginning, screen by screen, until no more new patterns appear. In this project, only 30 sentences were studied for each phrase, which is about the first two screens in Sinclair's view. By describing the translation units composed of the thirty A+N phrases and its translation equivalents, one is in a position to establish a sense of the properties and features of the translation units and their equivalence.

However, by using these thirty sample occurrences, there is a risk that not all types of translation units for a given A+N phase will be listed. Appendix 5 contains details of a more complete extraction approach which may be helpful for future studies. (Regrettably, time constraints meant that it was not adopted for this study.) The proposed method is more detailed in that it aims to extract all occurrences of each phrase and their translation equivalents.

The procedure was tested by extracting the translation equivalents of one of the selected A+N phrases – *medical officer*- and it appears that this method will lead to a more complete result than the thirty-sample extraction approach. Using the latter approach, only two translation equivalents for *medical officer* could be drawn: (1) 医生/yi sheng, (2) 公职医生 gong zhi yi sheng. However, by extracting and analysing all the occurrences (88) of the *medical officer*, a

third translation equivalent has been discovered: 医官/yi guan. This third equivalent only occurs three times in the whole corpus. In all three cases, *medical officer* always follows the word *chief*, and therefore they form a new translation unit – *chief medicinal officer* – and the translation equivalent of *chief medicinal officer* is 总医官/zhong yi guan.

This approach is equivalent to the way that one might order all the concordance lines of a node in the Bank of English in order to have an overview of all of its frequent patterns. The advantage of this approach is that nothing should be missed but the disadvantage is that it may be time-consuming if the concordance lines are numerous. Since insufficient time was available to do all of the complete extractions in this manner, the sample studies were conducted using the original sample extraction approach described above in Section 4.2.

4.4 Summary

In this chapter, I have introduced the approach taken in this analysis to extract thirty A+N phrases and their equivalents. This approach can be regarded as a mixture of automatic and manual procedures. Perl programs implement most steps in this approach but the extracting of translation equivalent is still dependent on manual work. The sample-extraction method is not completely satisfactory and a complete-search methodology has been tested since then which might be more useful to future researchers.

The Perl programme cannot extract the ready-to-use translation units but provides the candidates. Expert human interaction must be involved to finish the extraction process. The

problem of how to make the whole procedure fully automatic is probably a question yet to be resolved by workers in the field of computational linguistics.

Chapter 5

Translation Equivalents of A+N Phrases

A translation unit is defined as having only one meaning from the perspective of the target language. This meaning is believed to be represented as only one translation equivalent in the target language. However, not all the 30 sample phrases have only one equivalent. Since their equivalents have been manually identified and should be accurate, this phenomenon indicates that only some of the 30 A+N phrases are translation units while some are not. This chapter illustrates why some phrases have more than one equivalent. Section 5.1 provides an overview profile of the equivalence of all the 30 phrases; Section 5.2 discusses those phrases with only one translation equivalent; Section 5.3 analyses those with more than one translation equivalent.

5.1. The Overview of A+ N Translation Equivalents

5.1.1 The Occurrence of the Phrases in the Extracted Sentences

As has been mentioned in Chapter 4, 30 English A+N phrases were selected as the node, and their 30 aligned English and Chinese sentences were sampled as their contexts. If an A+N phrase occurs once in each of the English sentences, it will occur 30 times in the 30 extracted sentences. In some sentences, there may be more than one occurrence of an A+N phrase. Therefore, the total extracted occurrence is at least 30 times, but some phrases occur more than this in the 30 extracted sentences.

For example, there are 5 occurrences of *human remains* in the following extracted sentence pair (See Example 1) and all five occurrences have been translated into their translation equivalent 人类遗骸/*ren nei yi hai*:

Example 1

54679 Save in accordance with the provisions of this Part, any person who, without the permission in writing of the Authority, exhumes any [**human remains**] or any part of any **human remains** or any article interred therewith, or removes any **human remains**, or any part of any **human remains**, or any article from any urn or other receptacle, or removes or carries away any urn or other receptacle containing any **human remains** from any place, shall be guilty of an offence.

54679 除按照本部条文规定外，任何人在没有主管当局书面准许下，检掘**人类遗骸**、**人类遗骸**的任何部分或任何陪葬物品，从瓮盎或其它盛器移走**人类遗骸**、**人类遗骸**的任何部分或任何物品，或从任何地方移走或带走载有**人类遗骸**的瓮盎或其它盛器，即属犯罪。

The square brackets around the first *human remains* indicate that the phrase occurs for the first time here in this sentence. These brackets are added by the Perl program when the phrase is extracted. The brackets are used for convenience in locating the phrase. The following occurrences are indicated by the bold font. The translation keeps the original sequence of the five phrases. As a result of its multi-occurrence in some sentences, the phrase *human remains* occurs 42 times (both in English and Chinese) in the 30 extracted sentence pairs.

The A+N phrases are seldom omitted during the rendition. The nominal English phrases are generally translated into nominal Chinese phrases. Thus all the occurrences of the 30 English A+N phrases have been translated into their Chinese equivalents. The only exception is the phrase *straight line* in the following sentence:

Example 2

56797 Kwai Chung or Tsuen Wan bays, being all those waters between Tsing Yi Island and the mainland
1 , 2
bounded by a line drawn north from the northern extremity of Tsing Yi Island, a line drawn west from the
3 , 4
southern quivalen of Pillar Island, a line drawn from the southern extremity of Pillar Island to the western
4 , 5

extremity of Stonecutters Island and a [straight line] drawn true north from the westernmost extremity of
5 6 7 8 9
Stonecutters Island to the mainland
9 10
56797 葵涌海湾或荃湾海湾, 即青衣岛与大陆之间的所有水域, 界线范围为由青衣岛的最北端向北而
1 2 3
划, 由青洲的最南端向西而划, 由青洲的最南端划向昂船洲的最西端, 以及由昂船洲的最西端向正北
3, 4 5 6 9 8
而划向大陆。
7 10 .

The numbers under the above sentence show which parts of English fragments have been translated into their correspondent numbered fragments in Chinese. In the above sentence, the phrase *straight line* has been omitted during the rendition. The loss of the phrase in Chinese is because of a structural difference between Chinese and English. In Chinese, noun phrases can more often be omitted after transitive verbs if the context makes very clear to which noun phrases the verb refers. For instance, if a mother asks her little daughter to wash her face, and several minutes later, she wants to check if the child has done it or not, she may use the sentence “Have you washed?” instead of “Have you washed your face?” in Chinese. In the same way, an English mother might ask “Have you done it?” or even “Well, have you?” In translating the legal phrase given above, English requires the noun phrase to be repeated, but Chinese does not. This is what has happened in the translation of sentence 56797. The phrase *draw a straight line*/划直线 in English sentence has been reduced to *draw*/划 in the Chinese sentence. However, this kind of omission happens more often in colloquial Chinese than in written Chinese; this is why there is only one case of omission of the lexical equivalence in all these extracted phrases. Apart from this single case, other A+N phrases have all been translated into their Chinese equivalents.

5.1.2 The Frequency Calculation of the Phrases and Their Translation

Equivalents

Irrespective of whether a certain phrase occurs once or many times in a sentence, the relationship of the A+N phrase and that of its translation equivalent is usually one-to one, as in *Example 1* (every occurrence of *human remains* has its equivalent 人类遗骸). In very few cases, the relationship is many-to-one, or, less often, one-to-many. Take the phrase *personal representatives*; *Example 3* has more occurrences of the Chinese translation equivalent in the sentence pair while *Example 4* has more occurrences of the phrase itself.

Example 3

4403 When funds in court are by an order directed to be paid, transferred or delivered to any persons as legal [personal representatives], such funds or any portion thereof for the time being remaining unpaid, untransferred or undelivered may, upon proof of the death of any of **such representatives**, whether on or after the date of such order be paid, transferred or delivered to the survivors of **them**.

4403 如藉命令指示，法院储存金须支付、转拨或交付予任何身为合法**遗产代理人**的人，而该储存金或其中任何部分当其时仍未支付、转拨或交付，则一经证明任何一名该等合法**遗产代理人**已去世，该储存金或该部分可支付、转拨或交付予尚存的合法**遗产代理人**。

In the above example, there is only one occurrence of *personal representatives*. However, its Chinese equivalence 遗产代理人/ yi chan dai li ren occurs three times in the Chinese sentence. The relationship between the occurrence of the phrase and its translation equivalent is one-to-many. The second 遗产代理人/ yi chan dai li ren is rendered from *such representatives* which refers to the previous *personal representatives*. The third 遗产代理人/ yi chan dai li ren has been rendered from the pronoun *them*, which refers to the *personal representatives* as well. The grammatical words *such*, *which* and other pronouns (such as *them*) usually have an anaphoric function, i.e., they refer to the previous nouns for which they stand. When the frequency of the phrase and its equivalent are calculated, only the lexical phrases will be counted and the anaphoric variations will be ignored. This means, for the sentence above, the phrase *personal*

representatives is only calculated as occurring once, and its equivalent is also calculated as occurring once. That is, only those translation equivalents are calculated which are actually translated from the lexical phrases, but not those which occur due to additional translation techniques.

The following is an example of a many-to-one relationship between the occurrence of the English phrase *personal representatives* and its Chinese equivalent:

Example 4

13345 With a view to the conveyance to or distribution among the persons entitled to any movable or immovable property, trustees or **[personal representatives]** may give notice by advertisement in the Gazette, and such other like notices, including notices elsewhere than in Hong Kong, as would, in any special case, have been directed by a court of competent jurisdiction in an action for administration, of their intention to make such conveyance or distribution as aforesaid, and requiring any person interested to send to the trustees or **personal representatives** within the time, not being less than 2 months, fixed in the notice or, where more than one notice is given, in the last of the notices, particulars of his claim in respect of the property or any part thereof to which the notice relates.

13345 为向有权享有任何动产或不动产的人作出转易或分配, 受托人或**遗产代理人**可藉在宪报刊登公告而给予通知, 以及可给予具司法管辖权的法院在遗产管理诉讼的任何特殊个案中指示给予的其它同类通知, 包括在香港以外地方的通知, 以表明他们有意作出上述转易或分配, 并要求任何享有益的人, 在该通知所指定的不少于 2 个月的时间内, 或在该通知多于一份的情况下, 则在最后一份通知所指定的不少于 2 个月的时间内, 将他就该通知所关涉的财产或其任何部分而提出的申索详情, 送交**他们**。

In sentence 13345, *personal representatives* occurs twice in the English sentence, but there is only one translation equivalent of 遗产代理人/yi chan dai li ren in the Chinese counterpart.

The relationship is two-to-one. The first *personal representatives* has been translated as 遗产代理人/yi chan dai li ren, but the second *personal representatives* has been translated as a pronoun 他们/ta men (which means *them* in English) to avoid the repetition of the same noun phrase. In other words, the lexical phrase has been translated into the grammatical words — 他们/ta men. However, many-to-one cases are very rare (only one case in all the extracted

texts); in this dissertation, the calculation and analysis of translation equivalence will focus on the lexical translation equivalents. This means, the non-lexical equivalents will be ignored.

5.1.3 The Profile of the Chinese Translation Equivalence

As stated above, the hypothesis of this dissertation is that if a phrase, in this case an A+N phrase, is a translation unit, and therefore from the perspective of Chinese it has only one meaning, then this translation unit would normally have only one translation equivalent. While the extracted data confirm that in many cases there is only one translation equivalent, there are more in a number of cases.

Among all the 30 A+N phrases, there are 20 whose translation equivalents are the same. The rate of the same translation equivalence is over 60% (20 out of 30). These 20 phrases and their equivalents are listed in Table 2.

Table 2: 20 A+N phrases with one Chinese equivalence

A+N Phrase	Chinese Equivalent	Pin Yin of the Chinese Equivalent
annual allowance	年积金	nian ji jin
criminal offences	刑事罪行	xing shi zui xing
final appeal	终审	zhong shen
first column	第 1 栏	di yi lan
internal combustion	内燃	nei ran
judicial trustee	司法受托人	si fa shou tuo ren
legal adviser	法律顾问	fa lǚ gu wen
legal officer	律政人员	lǚ zheng ren yuan
listed company	上市公司	shang shi gong si
notifiable workplace	应呈报工场	ying cheng bao gong chang
personal representatives	遗产代理人	yi chan dai li ren
postal packet	邮包	you bao
provisional registration	临时注册	lin shi zhu ce
public bus	公共巴士	gong gong ba shi
rateable value	应课差饷租值	ying ke cha xiang zu zhi
registered dentist	注册牙医	zhu ce ya yi

registered scheme	注册计划	zhu ce ji hua
restricted licence	有限制牌照	you xian zhi pai zhao
special category	特种	te zhong
straight line	直线	zhi xian

The first column in Table 2 lists the 20 A+N phrases and the second column their translation equivalents. The third column is the Chinese Pinyin form of these translation equivalents in column 2. The occurrence, or frequency, of the phrases has not been listed because all their equivalents are the same, no matter whether they occur 30 or 40 times. Thus it is unnecessary to list their occurrences individually. In other words, the default frequency in these tables is all their occurrences in the extracted 30 sentences.

All these A+N phrases are translated into nominal phrases in Chinese. However, not all the 20 phrases belong to the same category if they are analysed in detail. Some of them are independent translation units while the others are not. In Section 5.2, I shall discuss these 20 phrases in detail.

The remaining 10 phrases have not been translated into the same Chinese equivalents. These 10 phrases are: *long term*, *light bus*, *conclusive evidence*, *written permission*, *good order*, *necessary modification*, *reasonable ground*, *medical officer*, *human remains*, and *residential care*. The phrase *good order* seems to have the most variable equivalents—it has five variations of Chinese equivalents in different contexts. Two phrases, *written permission* and *necessary modifications*, have four variations of Chinese equivalents. The remaining seven phrases have at least two translation equivalents. Details will be discussed in Section 5.3.

All the Chinese equivalents are nominal phrases except those for *long term* (see Section 5.3.2) and four of the equivalents of *good order* (see Section 5.3.3). Semantically, the translation equivalents of the 30 A+N phrases fall into 3 categories: 1) all the translation equivalents of an A+N phrase are the same; 2) the translation equivalents of an A+N phrase are not exactly the same, but they are synonymous; 3) the translation equivalent of an A+N phrase are neither the same nor synonymous. They are different in meaning. These three cases will be illustrated in the following sub-sections of this chapter.

5.2 The A+N Phrases with One Translation Equivalent

This section will analyse the 20 phrases listed in Table 2 which have one translation equivalent. Not all of these phrases are complete translation units. While some of them are independent translation units, some of them can be regarded as parts of larger translation units, and some can be both.

5.2.1 A+N phrases Functioning as Translation Units

The first type of phrase is those with one translation equivalent and they are complete translation units in all their occurrences. Both their meaning and their positions are independent in the sentences. To illustrate this type in detail, let us take *legal officer* for instance. Example 5 exemplifies the occurrence of *legal officer* and its translation equivalent in all the extracted sentences. In all the occurrences of the phrase, there is no other lexical word adjacent to it, neither to the left nor to the right.

Example 5

34527 “[**legal officer**]” means an officer appointed to and serving in the Colony as a **legal officer**, or an officer lawfully performing the functions of any of the officers designated in the Schedule;

34527 “**律政人员**”指获委任并在香港出任**律政人员**的人员，或合法执行附表内所指定任何人员的职能的人员；

The first *legal officer* occurs in quotation marks. It is like a proper noun, and its meaning does not depend on other words in the sentence. In other words, the text segment is disconnected from other parts of that sentence. The phrase itself is a unit of meaning. The situation of the second *legal officer* is the same, and the only difference is that the second one has a determiner which would usually be lost during the rendition (since Chinese does not have determiners like English). The translation equivalents reflect this analysis in the Chinese sentence. From the semantic point of view, the meaning of the phrase and its translation equivalence is independent from the further context. The phrase, therefore, is regarded as a translation unit in this dissertation.

There are 13 out of the above 20 phrases which can be regarded as whole translation units. These 13 phrases are shown in Table 3. They occur independently, without semantic interference from other pre-modified lexical words or post-modified lexical words. They do not have a strong collocability with other grammatical words either.

Table 3: 13 A+N Phrases with the same Translation Equivalents are also Translation Units

A+N phrase	Chinese Equivalence/Pinyin
straight line	直线/zhi xian
legal officer	律政人员/lü zheng ren yuan
criminal offences	刑事罪行/xing shi zui xing
annual allowance	年积金/nian ji jin
first column	第 1 栏/di yi lan
notifiable workplace	应呈报工场/ying cheng bao gong chang
listed company	上市公司/shang shi gong si
legal adviser	法律顾问/fa lü gu wen
registered dentist	注册牙医/zhu ce ya yi

postal packet	邮包/you bao
registered scheme	注册计划/zhu ce ji hua
judicial trustee	司法受托人/si fa shou tuo ren
rateable value	应课差饷租值/ying ke cha xiang zu zhi

In the initial hypothesis, it is assumed that each translation unit should have only one translation equivalent. These 13 phrases reflect the hypothesis. They are clearly translation units, and each of them has only one translation equivalent in all the extracted texts. These 13 phrases are regarded as non-problematic and will not be further discussed.

The focus will be on the other two types of phrases, both of which can be seen as part of a translation unit. They can be used independently, but they can also be extended left or right, and can be a part of a larger translation unit. The difference is that the second type occurs only as part of a larger translation unit, while the third type can occur both as a translation unit in its own right and as a part of a larger translation unit.

5.2.2 The A+N Phrases Functioning as Parts of Larger Units Only

The second type comprises those which are always part of larger translation units. This kind of phrase expresses independent meaning, but they are always followed by some other words, and form another larger unit of meaning throughout the corpus. For example, *special category* seems to be a translation unit, and its Chinese equivalent 特种/*te zhong* can be identified as well. However, in the corpus, it is always used together with another noun *space* -- either in singular form or plural form. This larger unit *special category space(s)* forms another meaning: 特种舱/*te zhong cang* in Chinese. Therefore, the phrase *special category* is not a complete translation unit, but only part of a complete translation unit *special category space(s)*. There

are altogether four A+N phrases which belong to this type and they have been displayed in Table 4. The larger units are listed in the middle column.

Table 4: A+N phrases as Parts of Larger Translation Units

A+N Phrase	Translation Unit	Chinese Equivalent/Pinyin
special category	special category space(s)	特种舱/te zhong cang
final appeal	(the) court of final appeal	终审法院/zhong shen fa yuan
restricted licence	restricted licence bank	有限制牌照银行/you xian zhi pai zhao yin hang
internal combustion	internal combustion engine/12	内燃机/nei ran ji/12
	internal combustion type machinery/8	内燃式机械/nei ran shi ji xie/8
	internal combustion marine machine/2	内燃船机/nei ran chuan ji/2
	internal combustion type propelling machinery/9	内燃式推进机械/nei ran shi tui jin ji xie/9

The Wordsmith software produces 94 concordance lines of *special category space(s)* when it runs the English part of the corpus. This is nearly the same frequency as the phrase *special category* extracted by the Perl program (92 times). This indicates that almost whenever *special category* occurs, it occurs with the word *space* either in singular or in plural form. This indicates that the phrase *special category* itself, in this corpus, is not an independent unit but normally requires the company of the third lexical word in order to make a full translation unit. All the instances of *special category space(s)* have been translated as 特种舱/te zhong cang in the Chinese text.

Similarly, *final appeal* does not occur alone but with *(the) court of final appeal*. This indicates that *final appeal* is only a part of a larger translation unit — *(the) court of final appeal*. In the translation, *(the) court of final appeal* has always been translated as 终审法院/zhong shen fa yuan.

The situation is the same for *restricted licence*. In the HKLDC parallel corpus, *restricted licence* always co-occurs with the word *bank* in order to make the whole unit of meaning. Whenever *restricted licence* occurs, it occurs as *restricted licence bank*. In this case, the patterned phrase (A+N) is only part of the larger unit. In the translation context, the larger units are the translation units. The Chinese equivalent of *restricted licence bank* is uniformly 有限制牌照银行/*you xian zhi pai zhao yin hang*.

The A+N phrase *internal combustion* is the same as the above three phrases except that it can be a part of more than one translation unit in this parallel corpus. It has formed four larger units: *internal combustion engine*, *internal combustion type machinery*, *internal combustion marine machine* and *internal combustion type propelling machinery*. The frequency of each of these units has been listed in Table 4. Each of the units have been translated as different Chinese phrases.

Although these four A+N phrases should belong to larger translation units in theory, their Chinese counterparts can be identified in the larger translation equivalents. That is, *special category* is 特种/*te zhong* in Chinese, *final appeal* is 终审/*zhong shen*, *restricted licence* is 有限制牌照/*you xian zhi pai zhao*, and *internal combustion* is always 内燃/*nei ran* no matter in which of the larger units it occurs. This may explain that why people have the false impression that texts are translated word by word. These complete translation units may not be identified by a reader who does not have an understanding of the structure of Chinese and English. In practice, the readers may feel satisfied if they are told the translation equivalents

of these A+N phrases. They may not be interested in what the whole translation unit is unless word to word matching is wrong.

Table 4 also shows that the syntactic pattern of a complete translation unit is not combined to bi-gram A+N. It may be tri-gram A+N+N (such as *special category space(s)*), or even n-grams (n>3) A+N+N+A+N (such as *internal combustion type propelling machinery*). There may also be a preposition in its pattern (such as *(the) court of final appeal*).

5.2.3 A+N Phrases Functioning both as Complete Translation Units and as Part of Larger Translation Units

The following phrases listed in Table 5 are another type of A+N phrase. They can either occur independently to form a translation unit, or they form another larger translation unit with adjacent words. Three A+N phrases belong to this kind: *personal representative*, *public bus*, and *provisional registration*.

Table 5: A+N Phrases Both as Translation Unit and as parts of Translation Units

A+N phrase	Translation Unit/Freq.	Chinese Equivalent/Pinyin/Freq.
personal representatives	personal representative/35	遗产代理人/yi chan dai li ren/35
	legal personal representative/4	合法遗产代理人/ he fa yi chan dai li ren/4
public bus	public bus/2	公共巴士/gong gong ba shi/2
	public bus service/30	公共巴士服务/gong gong ba shi fu wu/30
provisional registration	provisional registration/23	临时注册/lin shi zhu ce/23
	certificate of provisional registration/9	临时注册证明书/lin shi zhu ce zheng ming shu/9

There are 39 occurrences of the phrase *personal representative*. Among these, there are 35 occasions where *personal representative* occurs independently; that is, it occurs without the accompaniment of any other lexical words to form a unit of meaning. On all occasions

personal representative has been translated into 遗产代理人/*yi chan dai li ren*. Yet there are another four times where *personal representative* occurs with the word *legal* in the front to form another unit of meaning, *legal personal representative*. This new unit of meaning has been translated as 合法遗产代理人/*he fa yi chan dai li ren*. The word *legal* is polysemous in the English monolingual dictionary, and has more than one translation equivalent. According to *A New English-Chinese Dictionary*, *legal* has four Chinese equivalents: 法律的/*fa lü de*, 合法的/*he fa de*, 法定的/*fa ding de*, and 律师的/*lǚ shi de*. In the translation of *legal personal representative*, the word *legal* has lost the meaning of the other three translation equivalents and has been translated as 合法的/*he fa de*. The phrase *legal personal representative* should be identified as a translation unit because the whole phrase is unambiguous, while, from the Chinese perspective, *legal* has four meanings. The syntactic pattern of this translation unit is A+A+N—there is an adjective before the pattern A+N and it is different from the other patterns.

Like *personal representative*, *public bus* and *provisional registration* can also be extended into other translation units. The difference lies in the dominant phrases. For *personal representative* and *provisional registration*, the independent forms occur more often; *personal representative* occurs on 31 occasions more than *legal personal representative* in the total 39 occurrences; *provisional registration* occurs 14 occasions more than *certificate of provisional registration*. However, for *public bus*, the larger translation unit *public bus service* is more frequent (occurring on 28 occasions more than the independent form *public bus*).

The frequency of different forms may depend on the content of the text. Here, we would like to focus on the translation equivalents of these units. No matter whether they occur alone or as parts of larger units, these three A+N phrases have been translated as the same equivalents. To be specific, all occurrences of *personal representatives* have been translated as 遗产代理人/*yi chan dai li ren* whether they are used alone or in the larger unit *legal personal representative*. Similarly, all occurrences of *public bus* have been rendered as 公共巴士/*gong gong ba shi*, and all occurrences of *provisional registration* have been rendered as 临时注册/*lin shi zhu ce*.

They are also parts of larger translation units because together with the words added they form another meaning. The added lexical words, *legal*, *service* and *certificate* by themselves are polysemous from the Chinese perspective, and therefore not translation units. In this corpus, they become monosemous when they co-occur with *personal representatives*, *public bus* and *provisional registration*. A translator has to know that *service* in the phrase *public bus service*, is translated always as 服务/*fu wu*, never as 效力/*xiao li* or as 受雇/*shou gu*, the other alternatives given in the dictionary. The same is true for *certificate* in the phrase *certificate of provisional registration*.

5.3. A+N Phrases With More Than One Translation Equivalent

This section will analyse the other type of A+N phrases which have more than one translation equivalent. There are altogether 10 cases belonging to this type (see Section 5.1.3). They have more than one translation equivalent in the extracted 30 occurrences. In some cases, the

variations of translation equivalents are synonymous, in other cases they are not. Eight A+N phrases can be analysed as independent translation units. These 8 phrases can be further classified into two categories according to whether the variations are synonyms or not. Two phrases, *good order* and *residential care* cannot always be analysed as translation units. Other words must be searched for in order to form independent translation units. These complete translation units are then normally rendered by one translation equivalent. Section 5.3.1 will focus on the analysis of those translation units whose translation equivalents are synonymous, and Section 5.3.2 on the translation units whose equivalents are not synonymous. In Section 5.3.3, analysis will be focused on two special phrases, *good order* and *residential care*, which are parts of more complicated translation units.

5.3.1 A+N Phrases Whose Translation Equivalents are Synonymous

There are five A+N phrases listed in Table 6 having synonymous translation equivalents. These five phrases are regarded as complete translation units. Table 6 lists their synonymous translation equivalents. (TE in Table 6 is the abbreviation of Translation Equivalent.) The order of the TEs is based on their frequencies. That is, the most frequent equivalent is regarded as the first TE, and the immediate less frequent one is listed as the second and so on.

Table 6: 5 A+N Phrases Whose Translation Equivalents are Synonymous

A+N Phrase	1 st TE /Pinyin/Freq.	2 nd TE /Pinyin/Freq.	3 rd TE /Pinyin/Freq.	4 th TE/Freq.
light bus	小 巴 /xiao ba/31	小型巴士/xiao xing ba shi/22		
written permission	书面准许/shu mian zhun xu/17	书面许可/shu mian xu ke/7	书面批准/shu mian pi zhun/3	准许 /zhun xu/3
necessary modifications	必要的变通/bi yao de bian tong/20	必需的变通/bi xu de bian tong/7	需要的变通/xu yao de bian tong/2	必需的修改/bi xu de xiu gai/1
reasonable ground	合理的理由/he li de li you/16	合理理由/he li li you/15		

human remains	人类遗骸/ren nei yi hai/41	遗骸/yi hai/1		
------------------	---------------------------	-------------	--	--

Most of the five A+N phrases have only two translation equivalents, but *written permission* and *necessary modification* have four translation equivalents respectively. In both cases, their translation equivalents are synonymous and can replace each other in all contexts.

合理的理由/he li de li you and 合理理由/he li li you, the two translation equivalents of *reasonable ground*, are the two most obvious synonyms. Their difference is that the first translation equivalent has a Chinese character 的/de while the later one has not. The Chinese character 的/de in 合理的理由/he li de li you is used as an adjective suffix, which can be, and often is, omitted to achieve concision.

小巴/xiao ba and 小型巴士/xiao xing ba shi are both rendered from *light bus*. 小巴/xiao ba is an abbreviation form of 小型巴士/xiao xing ba shi. 小型/xiao xing has been abbreviated as 小/xiao and 巴士 ba shi has been abbreviated as 巴/ba. Although 小巴/xiao ba may be used more in spoken Chinese and 小型巴士/xiao xing ba shi sounds more formal, their referential meanings are the same. They are synonymous as well.

人类遗骸/ren nei yi hai and 遗骸/yi hai from *human remains* are not synonyms if we consider them as two separate terms. 遗骸/yi hai means *remains* and has broader meaning than *human remains*(人类遗骸/ren nei yi hai). 遗骸/yi hai can include not only the remains of human beings, but also the remains of animals, fish, plants and so on. Therefore, the first

glance may give a reader a false impression that these two translation equivalents are referentially different. There is, however, only one case of 遗骸/yi hai, but the rest are all rendered as 人类遗骸/ren nei yi hai. The context of this occurrence is given in *Example 6*:

Example 6:

54740 Where a person who has the right to effect the disposal of the **human remains** of any person-
54741 within the period of 48 hours after the **human remains** are received into any mortuary-

54740 如具有处置任何**人类遗骸**的权利的人—

54741 在验房接收**该遗骸**后 48 小时的期限内—

Sentences 54740 and 54741 in fact belong to the same semantic sentence/clause, but they have been cut into two sentences for the sake of alignment during the corpus building. If we read them together as one single sentence, we may find that the two *human remains* actually refer to the same thing. Then why has the second *human remains* been translated differently? The secret lies in the Chinese character 该/gai before 遗骸/yi hai in sentence 54741. 该/gai is an anaphor and means *such* in Chinese. 该遗骸/gai yi hai means *such remains*, which refers to the same human remains mentioned before. In fact, in this case, 遗骸/yi hai and 人类遗骸/ren nei yi hai share the same referential meaning because of the Chinese functional character 该/gai. Therefore, the whole translation equivalent is not 遗骸/yi hai but 该遗骸/gai yi hai. 该遗骸/gai yi hai and 人类遗骸/ren nei yi hai are synonymous, sharing the same referential meaning.

Although *written permission* and *necessary modifications* seem to have more equivalent variations, their translation equivalents are synonymous. For 准许/zhun xu, 许可/xu ke, and 批准/pi zhun are synonyms, which means permission or giving permission. In the first three TE, the word *written* has been rendered as 书面/shu mian. The first three equivalents 书面准

许/shu mian zhun xu, 书面许可/shu mian xu ke and 书面批准/shu mian pi zhun are synonymous. 书面准许/shu mian zhun xu and the fourth translation equivalent 准许/zhun xu fall into the same category of 小巴/xiao ba and 小型巴士/xiao xing ba shi. 书面/shu mian is omitted for the sake of conciseness.

In the four translation equivalents of *necessary modifications*, *modification* has nearly always been translated as the same 变通/bian tong except in one sentence as 修改/xiu gai. The three variations translated from *necessary* — 必要的/bi yao de, 必需的/bi xu de and 需要的/xu yao de — are synonymous in Chinese. Therefore, the first three translation equivalents are synonymous. Since 修改/xiu gai and 变通/bian tong are synonymous as well, the fourth translation are synonymous with the previous three.

5.3.2 A+N Phrases Whose Translation Equivalents are not Synonymous

There are five A+N phrases which have non-synonymous translation equivalents. Two A+N phrases, namely, *good order* and *residential care*, are more complicated; and they will be analysed in the next section (5.3.3). The remaining three phrases and their equivalents are listed in Table 7.

Table 7: 3 A+N Phrases Whose Translation Equivalents are not Synonymous:

A+N Phrase	1 st TE/Pinyin/Freq.	2 nd TE/Pinyin/Freq.
long term	长远/chang yuan/36	长期/chang qi/2
conclusive evidence	确证/que zheng/27	不可推翻的证据/bu quiva fan de zheng ju/5
medical officer	公职医生 gong zhi yi sheng/18	医生/yi sheng/14

The two translation variations of *long term* are due to their different contexts. In fact, *long term* belongs to another two translation units – *long term business* and *long term interest*. 长远/*chang yuan* and 长期/*chang qi* may be synonymous in other cases, but here they cannot replace each other, therefore they are regarded as non-synonymous. When it is used in *long term interest*, *long term* is always translated as 长远/*chang yuan*; however, it is translated as 长期/*chang qi* in the larger translation unit *long term business*. This shows that *long term* itself is not an independent translation unit, for it has to be accompanied by *business* or *interest* to form a unit. These two translation units are monosemous from the Chinese perspective.

The two translation versions of *conclusive evidence* are to some degree synonymous, but they cannot strictly be called synonyms. Native Chinese speakers will understand what is meant by 确证/*que zheng* (back translation, *factual evidence*) and 不可推翻的证据/*bu ke tui fan de zheng ju* (back translation, *evidence impossible to overthrow*). The two translation alternatives are similar in that in both cases the evidence exists or they are referring to a fact. The difference is that the former Chinese translation focuses on the evidence, while the latter emphasises the impossibility of overthrowing the evidence. This difference is not caused by the inconsistency of the translation but by the slight difference of contexts.

Appendix 6 gives ten sample extracted sentences. In six of the ten sentences, *conclusive evidence* has been translated as 确证/*que zheng*, while in the remaining four sentences, it has been rendered as 不可推翻的证据/*bu ke tui fan de zheng ju*. The software did not identify much difference between their first 10 collocates within 5R-5L, which are mainly

grammatical words. If we take a close look at the contexts we may find similarities within the four sentences which have been translated as 不可推翻的证据/*bu ke tui fan de zheng ju*. Their context words include words implying criminal justice: *offence*, *prejudice*, *proceedings*, *prejudicial*, and *criminal*. However, in the six sentences which have been translated as 确证/*que zheng*, there are no such words. The sentences deal with issues connected with civil law. The different translations reflect the difference of their contexts. Thus *conclusive evidence* occurs in two different domains and forms two different translation units: one is with the criminal justice words while the other is not. This knowledge of domains is needed to disambiguate between the two meanings.

The two translations of *medical officer*, 公职医生 *gong zhi yi sheng* and 医生/*yi sheng*, refer to two different kinds of doctors. The same English phrase *medical officer* means different kinds of things in different legal documents as shown in Example 7:

Example 7

A:

62026 “[*medical officer*]” means a registered medical practitioner in the full time employment of Government or the Hospital Authority within the meaning of the Hospital Authority Ordinance;

62026 “公职医生” 指全职受雇于政府或全职受雇于《医院管理局条例》所指的医院管理局的注册医生;

B:

46562 “*Medical Officer*” means a Government [*medical officer*] assigned to a detention centre by the Director of Health or a doctor assigned to a detention centre by a voluntary agency as approved by the Secretary for Security;

46562 “医生” 指由生署署长派驻羁留中心的政府医生, 或由保安司认可的志愿机构派驻羁留中心的医生。

In Sentence pair A, *medical officer* is translated as 公职医生 *gong zhi yi sheng*, while in B, the same phrase has been translated differently as 医生/*yi sheng*. Both of these two *medical officers* occur at the beginning of their sentences, and both of them have been surrounded by

quotation marks. However, the same term is used to refer to two different concepts. 公职医生 *gong zhi yi sheng* is a full time government medical practitioner, regulated by the *Hospital Authority Ordinance*. 医生/*yi sheng*, however, is a government doctor assigned to a detention centre.

The translation of 公职医生/*gong zhi yi sheng* occurs in Chapter 136 2(1), while translation equivalent 医生/*yi sheng* occurs in 298A 2. The two documents are parts of different laws. Chapter 136 2(1) is the interpretation part of the MENTAL HEALTH ORDINANCE which was issued on 1st February, 1999. Chapter 298A 2 is the interpretation part of the PROBATION OF OFFENDERS RULES., which was issued on 30 June, 1997. In the English versions of these two different laws, the same *medical officer* refers to different concepts. When they are translated into Chinese, the translators purposely choose different Chinese terms to indicate this difference. The phrase *medical officer* forms two translation units when it appears in these two laws.

The following table lists the most frequent lexical words and grammatical words which co-occur with the two senses respectively. These words are limited within 5L to 5R:

Table 8: Collocates of *medical officer*

Frequency Rank	5 Most Frequent Lexical Collocates (5L -5R)		10 Most Frequent Grammatical Collocates (5L – 5R)	
	<i>medical officer</i> as 医生/ <i>yi sheng</i>	<i>medical officer</i> as 公职医生/ <i>gong zhi yi sheng</i>	<i>medical officer</i> as 医生/ <i>yi sheng</i>	<i>medical officer</i> as 公职医生/ <i>gong zhi yi sheng</i>
1	government	charge	the	the
2	means	mental	a	of
3	appointed	practitioner	of	a
4	charge	prison	shall	in
5	report	be	to	or
6			by	by

7			<i>or</i>	that
8			<i>in</i>	<i>to</i>
9			on	<i>any</i>
10			<i>any</i>	such

The same collocates are highlighted as italic bold words. It can be seen that the two senses of *medical officer* have not much difference in their grammatical collocates, but their lexical collocates are very different. Therefore, the two senses can be distinguished according to their most frequent context words. This example shows that some translation units can be extended to the register or the whole domain.

5.3.3 Two Special A+N Phrases as Parts of Translation Units

Good order and *residential care* share two similarities. They have more complicated translation equivalents and form more than one complete translation unit. Their translation equivalents cannot be identified without considering the whole translation unit. There is no word-for-word equivalence that can be identified here. To illustrate it clearly, let us look at the concordance of the 30 extracted examples of *good order*:

Figure 1: Concordance of *good order*:

1 60466 the maintenance of decency and [good order] in the stadium is prejudice
2 ner. 44679 maintenance of peace and [good order] in any place licensed under
3 s; 54311 maintenance of peace and [good order] in any place licensed under
4 ered, drained, lighted or maintained in [good order], the Building Authority-
5 sanitary condition and shall be kept in [good order] and repair. 56714 Every
6 g Authority, and shall be maintained in [good order] to his satisfaction, by the
7 nd sanitary condition and to be kept in [good order] and repair. 56977 Every
8 articles have been delivered but not in [good order] and condition, of the quiva
9 in a clean condition and maintained in [good order] and repair. 57115 Every
10 in a clean condition and maintained in [good order] and repair. 58655 Every
11 icer, and shall deliver the articles in [good order] and condition, fair wear an
12 tion or of maintaining such shoring in [good order] or of inspecting the same.
13 keep a public dance hall shall maintain [good order] in the premises and shall n
14 to keep a dancing school shall maintain [good order] in the premises and shall n
15- 58752 The licensee shall maintain [good order] on the licensed premises an
16- 58693 The licensee shall maintain [good order] on the licensed premises an
17 any stadium; 54566 preservation of [good order] and prevention of abuses an
18 he notice: 54111 the maintenance of [good order] in slaughterhouses; 5
19 nuisances; 54733 the maintenance of [good order] in public funeral halls.
20 ts of a detainee or in the interests of [good order] in the Centre that a detain
21 his Part; 54434 the preservation of [good order] and discipline and preventi
22 shall not interfere with the running or [good order] of the centre and is otherw

23 terest on the grounds of public safety, [good order] and security, the cost of t
 24 n an offensive trade to be kept in such [good order], repair and condition as to
 25 be kept clean and shall be kept in such [good order], repair and condition as to
 26 be kept clean and shall be kept in such [good order], repair and condition as to
 27 noxious matters, and to be kept in such [good order], repair and condition as to
 28 noxious matters and to be kept in such [good order], repair and condition as to
 29 ion on any problem which may affect the [good order] or discipline of the centre
 30 person to do any act prejudicial to the [good order] and security of the centre.

According to the context in Figure 1, *good order* can have 3 different senses:

1) *good order* is used to mean the good discipline of a place or premises. In this sense, if a verb such as *maintain* or *keep* and *affect* is used before it, *good order* is translated as 良好秩序 /*liang hao zhi xu* (1,2,3, 13, 14, 15, 16, 20, 22, 23, 29, and 30). If we find a noun rather than a verb before it, such as *maintenance* or *preservation*, *good order* is translated as 秩序良好 /*zhi xu liang hao* (17, 18, 19, 21).

2) *good order* is used with *maintain* or *keep* to refer to the status of an object. If the words following it are *repair* or *condition*, *good order*, together with the verb, is translated as 保持完好 /*bao chi wan hao* (5, 7, 9, 10, 24, 25, 26, 27 and 28). Without the words of *repair* or *condition* following it, *good order* is translated as 妥善 /*tuo shan* (6, 8, and 14).

3). *Good order* also means the property and sequence of certain articles. Usually, the preceding verb is *deliver*. It is translated as 性能良好 /*xing neng liang hao* (10 and 13).

We find that there are three meanings of the phrase *good order* in this case but they yield five translation units with their respective translation equivalents. All these extended translation

units are shown in Table 9. Among their five Chinese translation equivalents, only the first one is a nominal phrase. The second is an adjective phrase and the others are verbal phrases. This example gives some indication that Tufis (2001)'s assumption that translation units tend to be translated into the same syntactic category in the target language is not always true. This also explains why some algorithms based on this assumption will not gain the high precision.

Table 9: Translation Equivalents of *good order*

A+N Phrase	Whole Translation Unit	Chinese Equivalents/Pinyin/Freq.
good order	(keep/maintain)... good order (in some place)/12	(保持某处)...良好秩序/(bao chi mou chu) liang hao zhi xu/12
	(maintenance/preservation of) good order (in some place)/4	(保持某处)...秩序良好/(bao chi mou chu) zhi xu liang hao/5
	(something to be kept /maintained... in) good order (repair or condition)/9	(某物被保持)完好/(mou wu bei bao chi) wan hao/9
	(maintain) in good order/3	妥善(保养)/tuo shan (bao yang)/3
	(be delivered in) good order (and condition)/2	(保持)性能(和状况)良好/(bao chi) xing neng (he zhuang kuang) liang hao/2

With regard to the analysis of the phrase *residential care*, Figure 2 is the concordance of this term extracted from the corpus.

Figure 2: Concordance of *residential care*

```

1 rel home for elderly persons.      171168 "residential care home" means any pre
2 e to which section 33 applies.      43057 "residential care home" means any pre
3 rel home" means any premises-      43062 "residential care expenses" means any
4 s are necessary for the inspection of a residential care home or for the inspe
5 n of the licence issued in respect of a residential care home; and      171271 d
6 or has not been issued in respect of a residential care home shall be evidenc
7 person acting on its behalf;      51066 a residential care home for elderly pers
8 erson holding a licence in respect of a residential care home cancel or quival
9 eping, management or other control of a residential care home, as he thinks fi
10 erson holding a licence in respect of a residential care home may, before the
11 art in the operation or management of a residential care home to produce any b
12 ertificate of exemption in respect of a residential care home shall be-      171
13 , manage or otherwise have control of a residential care home of a type prescr
14 [residential care] homes;      171177 any residential care home used or intended
15 052 Where a deduction in respect of any residential care expenses is claimed b
16 ragraph , a deduction in respect of any residential care expenses shall not be
17 272 The Director may, in respect of any residential care home, by notice in wr
18 pays during any year of assessment any residential care expenses in respect o
19 uiring medical treatment;      171178 any residential care home or type or quiva
20 dparent of the person, in so far as any residential care expenses described in
21 reasonable times enter and inspect any residential care home or any premises
22 ration Ordinance ,to be an inspector of residential care homes.      171268 at a
23 nder section 17 to be an inspector of residential care homes;      171177 any
24n indictable offence in respect of that residential care home;      171215 on th
25n application for the licensing of that residential care home;      171214 any o
26 on the ground that, in respect of that residential care home or the residents

```

27 keep, manage or otherwise control, that residential care home; or 171222 th
 28 [residential care] home; 171220 that residential care home has ceased to be
 29 sidential care] home; or 171222 that residential care home has, on any occa
 30 n holding the licence in respect of the residential care home; 171220 that
 31 ing, management or other control of the residential care home, in addition to
 32 n holding the licence in respect of the residential care home; 171218 on th
 33 allowance included, in the case of any residential care expenses so claimed, a
 34 tion 33 applies and, in the case of any residential care expenses so allowed, a
 35 e Scheme, a deduction in respect of the residential care expenses shall be allow
 36 t of the residential care received at a residential care home and paid to that r
 37 al care] home or type or description of residential care home excluded by the Di
 38 residential care home and paid to that residential care home or any other perso
 39 g to the operation or management of the residential care home or to any other ac
 40 to any other activity in respect of the residential care home, or to furnish any
 41 eping, management or other control of a residential care home. 171272 The Dir
 42 ct are used as or for the purposes of a residential care home; 171269 require
 43 any expenses payable in respect of the residential care received at a residenti

In the extracted 30 sentences, there are 43 occurrences of *residential care*. Among all the 43 cases, there is only once case (line 43) where *residential care* is used alone. *Residential care* here has been rendered as 住宿照顾/*zhu su zhao gu*. There are 8 cases (3, 15, 16, 18, 20, 33, 34, 35) where *residential care* is used together with *expenses* and thus forms a larger translation unit of *residential care expenses*. *Residential care expenses* has been translated as 住宿照顾开支/*zhu su zhao gu kai zhi*. In the above 9 cases, whether it has been used alone or together with *expenses*, *residential care* has been translated as 住宿照顾/*zhu su zhao gu*. For the remaining 34 cases, *residential care* is part of the unit *residential care home*. In these 34 cases, *residential care* has lost its meaning as 住宿照顾/*zhu su zhao gu*, and the whole term (*residential care home*) has been translated as 安老院/*an lao yuan*. That is, 安老院/*an lao yuan* is the translation equivalent of the whole translation unit *residential care home*. The three Chinese characters do not match the three English words one-to-one, but the whole English term matches the whole Chinese term. The meaning of *residential care* in the phrase *residential care home* cannot be separated. This is a good example of unit-to-unit but not word-to-word translation. If a translator translates word-to-word here, the result will be unidiomatic translation.

Thus, there are three translation units: *residential care*, *residential care expenses*, and *residential care home*. This is shown in Table 10. Again, for the latter two translation units, their patterns are no longer A+N, but A+N+N.

Table 10: Translation Equivalents of *residential care*:

A+N Phrase	Whole Translation Unit	Chinese Equivalents/Pinyin/Freq.
residential care	residential care/1	住宿照顾/zhu su zhao gu/1
	residential care expenses/8	住宿照顾开支/zhu su zhao gu kai zhi/8
	residential care home/34	安老院/an lao yuan/34

5.4 Conclusion

In this section, I have analysed the translation equivalents of the extracted 30 A+N phrases. Only lexical equivalents are considered in this analysis. Starting from the phrases which have the same translation equivalent, I have illustrated all these phrases with only one translation equivalent if they are expanded into a complete translation unit. An A+N phrase is not necessarily a complete translation unit. The typical example is the translation of *residential care home*. The most complicated expansion of a translation unit is *good order* which has been expanded into five translation units. The context may not just be the adjacent words to the left or right, but may include words at some distance from the phrases. Sometimes even the whole domain will be the factor which helps to disambiguate the meaning, like *medical officer*; therefore, the different domain should be included in the complete translation units as well.

Not all A+N have been translated as the same A+N syntactically in Chinese (e.g. *good order*). However, each complete English translation unit has one translation equivalent in Chinese. If

for any reason they have more than one translation equivalent, these equivalents are synonymous. This supports the hypothesis: if the translation is consistent, a translation unit has only one translation equivalent; if a translation unit has more than one equivalent, they are synonymous. If a translation unit candidate has more than one non-synonymous equivalent, this candidate belongs to different complete translation units.

Chapter 6

Comparisons with Bilingual Dictionaries

The previous chapter contained an analysis of all the 30 A+N phrases and their translation equivalents. Some of them are complete translation units while some are parts of complete translation units, and some can be both. The analysis shows that an A+N phrase will have a single translation equivalent or synonymous equivalents if it is a translation unit; if an A+N phrase has more than one translation equivalent, it usually belongs to different translation units.

This chapter will compare the translation equivalents of the 30 A+N phrases provided by the HKLDC parallel corpus with those arrived at by consulting bilingual dictionaries. The two dictionaries used in the comparison work are: *A New English-Chinese Dictionary (Century edition)* (NECD for short) and *English-Chinese Glossary of Legal Terms (Web version)* (ECGLT for short). The former is a general bilingual dictionary widely used in Mainland China; the latter is a specialized bilingual dictionary published in Hong Kong. Through comparisons with these dictionaries, the deficiencies of traditional dictionaries will be shown when used as an aid for translating into a language that is not the native language of the dictionary-user; i.e. a language in which the user is not completely competent. The comparisons focus on two points. First, whether the corpus-derived English A+N phrases are listed in the English-Chinese dictionaries. If they are listed, are their Chinese translations the same as the corpus-derived translation equivalents? Second, if an English A+N phrase is **not**

found in the dictionary, whether a combination of the dictionary translation of the A part of the phrase and the dictionary translation of the N part of the phrase achieves an equivalent translation to the corpus-derived translation equivalent for that A + N phrase. This analysis is based on the hypothesis that using the translation units and their translation equivalents extracted from parallel corpora can facilitate translations into a non-native language and that traditional bilingual dictionaries frequently fail their users because they record these full translation units only sporadically.

The remainder of this chapter is organised as follows: Section 6.2 gives a brief introduction to the two dictionaries used in the comparison; Section 6.3 compares the corpus equivalents of the 30 A+N phrases with those of the NECD; Section 6.4 presents a comparison with ECGLT; and Section 6.5 concludes this chapter.

6.1 Two Dictionaries Used in the Comparison

6.1.1 Introduction of a New English-Chinese Dictionary (Century Edition)

The NECD is a general English-Chinese bilingual dictionary. It is one of the most popular dictionaries amongst English learners and users in Mainland China. Many distinguished Chinese experts in bilingual lexicography, English-Chinese language teaching and translation were involved in its compilation. Altogether, it has gone through three editions, and the most up-to-date version, the Century Edition, was published in 2000. For the Century Edition, many British and American English dictionaries published in the 1990s, both general and specific, were consulted. According to its Preface, it has more than 100,000 entries, including

recently adopted English words, new senses and new usages, which reflect the changes since the 1970s in societies, technologies, economies and cultures.

This dictionary was published after the documents in the HKLDC had been translated and issued. Since it is considered to provide the most updated definitions of the words and include the latest terms, it offers the best chance of finding all the words extracted from the HKLDC.

Because this general bilingual dictionary is published by a mainland press — Shanghai Translation Publishing House — the translation equivalents listed in the dictionary are those used on the mainland; these are sometimes slightly different from the equivalents used in Hong Kong. Also, the translation equivalents of daily words such as *light bus* (= mini-bus) are different from those used on the mainland. However, the legal language in general, and legal terminology in particular, with their translation equivalents, to the extent they have been included, should be similar or identical to those found in Hong Kong legal documents. This is not only because the Chinese government has applied the same standard of legal language and terminology both in Mainland China and Hong Kong, but also because the translation work has been done by a committee which is mostly made up of mainland legal experts and translators. (There were eight mainland members and four Hong Kong members in the Hong Kong Legal Document Draft Committee).

Two specific goals were pursued in consulting the general dictionary. First, to compare the differences between the translation equivalents in the parallel corpus and those arrived at by consulting the dictionary; second, to compare mainland and Hong Kong Chinese.

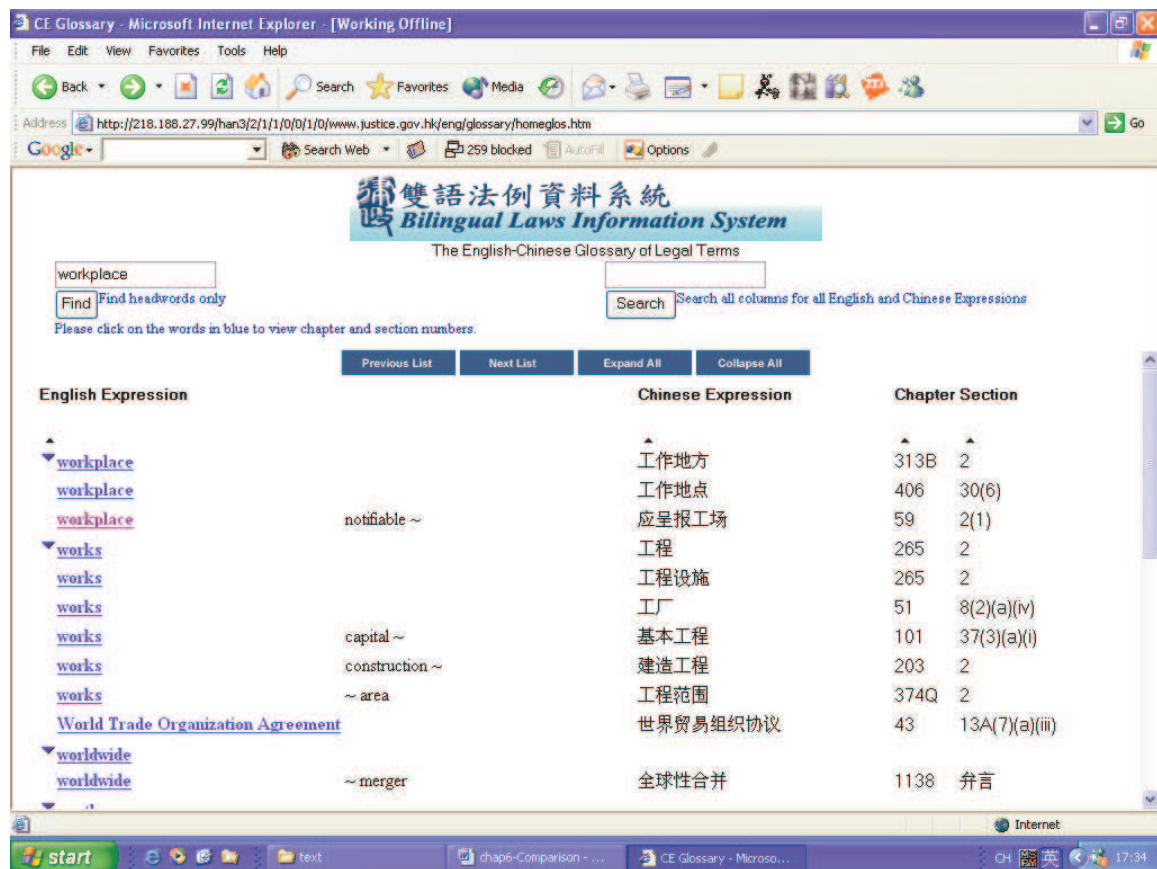
Not all the A+N phrases extracted are legal terms, but still they need to be translated correctly. The general dictionary should provide information on how these general terms would be translated. Since the legal documents are all from Hong Kong, it is also useful to establish to what extent the Chinese equivalents extracted from the corpus are different from the ones offered by the dictionary.

6.1.2 Introduction of English-Chinese Glossary of Legal Terms (Web version)

The English-Chinese Glossary of Legal Terms is published by the Law Drafting Division of the [Department of Justice](#) in Hong Kong in order to promote bilingualism in legal matters. It was compiled from materials contained in bilingual legislation. It has three printed editions, those of March 1995, June 1996, and September 1998. The web version of the English-Chinese Glossary of Legal Terms (ECGLT) is provided by the Bilingual Laws Information System (BLIS), a searchable electronic database of the statute laws of Hong Kong established and updated by the Department of Justice of the HKSARG (The Government of Hong Kong Special Administrative Region of the People's Republic of China). It is available at:

http: www.justice.gov.hk/eng/glossary/homeglos.htm.

The web version is arranged in this manner:



Picture 1 : The Web Version of the English-Chinese Glossary of Legal Terms

The first column contains English head words; the second column provides the collocation of the head words; the third column shows the Chinese translation of the head word or the term (if applicable); the last two columns provide the source of this term or head word.

The page above is the search result using the headword *workplace*. The first column is the English headwords column, where we can see that *workplace* is listed as a headword three times. In the first two instances, *workplace* has been translated as 工作地方/*gong zuo di fang* and 工作地点/*gong zuo di dian* respectively, but in the third instance, *workplace* has a

collocate *notifiable*, and the term *notifiable workplace* is the equivalent of 应呈报工厂/*ying cheng bao gong chang* which is the same as that found in the HKLDC.

Apart from the fast search speed, the web version has two other advantages. First, it provides an ‘all columns’ search function (the right query square) as well as the head word search. Second, it generates a link to the source where the head word, or term, comes from. These two functions mean that the web version is significantly more useful than the printed version. This might suggest the future direction of dictionary development.

Clicking on a headword (such as the third headword *workplace* in this example) leads one to the context, which, in this case, is chapter 59, section 2(1) as indicated on the right of the screen.

The ‘all columns’ search is more powerful than the ‘head word’ search function. It searches all the columns of the ECGLT. The ‘all columns’ search result of *notifiable workplace* looks as follows:

workplace — [*notifiable workplace* 应呈报工场 59 2\(1\)](#)

The first word is the headword (in this case: *workplace*); and the text following the dash is the composite term which has been searched (*notifiable workplace*). There follows the Chinese equivalent of the term, and also the numbers of the chapter and section in which the term occurs (chapter 59 section 2(1)), it is directly linked to the original text as well.

The web legal glossary ECGLT is linked to the database which comprises all the bilingual legal documents from Hong Kong, including the texts in the parallel corpus HKLDC. Therefore, it seems probable that the Chinese expressions in the ECGLT would be the same as those extracted from the HKLDC. However, the search result of these 30 A+N phrases does not meet this assumption. Some A+N phrases can be found in the ECGLT but their Chinese equivalents are not the dominant ones which the corpus would suggest; some phrases are listed in ECGLT but under unpredictable headwords. Moreover, some phrases have even been excluded in ECGLT altogether. Further details will be discussed in Section 6.4.

6.2 A Comparison of the Corpus with the New English-Chinese Dictionary (NECD)

Like all traditional dictionaries, NECD regards the single word as the default unit of analysis. Since many single words are ambiguous, there is normally more than one translation equivalent under a headword in the bilingual dictionary. For an A+N phrase, there are two different cases: (1) the phrase is found in the NECD as a subentry or an example under a headword; (2) the phrase is not listed in the dictionary. In the former case, there is no trouble making a comparison, but in the latter case, a problem is encountered; namely, how to find the meaning of the phrase in the dictionary.

To solve the above problem, a new idea is introduced as the solution and this is the concept of a phrase's default translation. In this dissertation, the default translation of an A+N phrase is defined as a combination of the default meaning of each word in the phrase. The default

meaning of a word is the first translation equivalent provided by a bilingual dictionary. Take the phrase *straight line* for example. Assume the first translation of *straight* in the dictionary is 直的/*zhi de* and the first translation of *line* is 线/*xian*. Then the default translation of the phrase *straight line* is 直的线/*zhi de xian*. Such a translation is called the default translation of a phrase.

In a comparison between the corpus and the dictionary, the default translation of a phrase is applied if the whole phrase is not listed in the NECD. This approach is adopted because non-native speakers, if they cannot find the exact phrase in the dictionary, may choose the default translation of this phrase as its translation equivalent. If a phrase, for any reason, has more than one translation equivalent in the corpus, the most frequent one is used to compare it with the default translation offered by the dictionary. The most frequent equivalent is called the corpus dominant translation in this dissertation.

When a comparison is made with the NECD, three phrases amongst the 30 A+N phrases are found in the NECD as subentries and another three phrases as examples given in their headword entries. These subentries and examples indicate that the lexicographers have noticed the phenomenon that these words co-select with each other and so have treated them as distinct terms. The three phrases, which are subentries in the dictionary, are listed in Table 11.

Table 11: Three A+N Phrases Listed as Subentries in the NECD

A+N Phrase	Dictionary Default Translation	Corpus Dominant Translation
long- term	长期的/ <i>chang qi</i>	长远/ <i>chang yuan</i> /36
internal combustion (engine)	内燃(机)/ <i>nei ran (ji)</i>	内燃(机) / <i>nei ran (ji)</i>
medical officer	卫生官员/ <i>wei sheng guan yuan</i>	公职医生 <i>gong zhi yi sheng</i> /18

In the NECD, *long term* has been hyphenated as *long-term*. In Table 11, *long-term* and *medical officer* are listed as subentries in the NECD under the headwords of *long* and *medical* respectively, while *internal combustion engine*, but not *internal combustion*, is a subentry in the dictionary.

Intuitively, the Chinese translation of these phrases should be the same or approximately the same as the corpus translation because they are regarded as fixed terms. However, the comparison shows that only one out of three dictionary translation equivalents is the same as that in the corpus. The sample study in the last chapter yields, in the corpus, two translation equivalents of *long term* — *长远/chang yuan* and *长期/chang qi*. In the corpus, *长远/chang yuan* is dominant, yet the dictionary provides only the less frequent one *长期的/chang qi de*. For *medical officer*, the corpus translations are *医生/yi sheng* and the less dominant one *公职医生/gong zhi yi sheng*. However, the dictionary provides neither of these but another one *卫生官员/wei sheng guan yuan*. While the equivalent of *卫生官员/wei sheng guan yuan* which it offers may be acceptable, it fails to provide those equivalents that are widely used in Hong Kong legal documents.

Table 12: Three A+N Phrases Listed as Examples in the Dictionary

A+N Phrase	Headword	Dictionary Default Translation	Corpus Dominant Translation
conclusive evidence	conclusive	确证/que zheng	确证/que zheng/27
listed company	list	上市公司/shang shi gong si	上市公司/ shang shi gong si
postal packet	packet	小件邮包/xiao jian you bao	邮包/ you bao

Table 12 shows three A+N phrases appearing in the NECD as examples, but not as subentries. They are, in the dictionary compilers' eyes, not sufficiently fixed as subentries, but the words in the phrases do appear together quite often in the experts' discourse. For example, *conclusive evidence* has been listed as an example under the headword entry *conclusive*; *listed*

company appears under the headword entry *list*; and *postal packet* under the headword entry *packet*. The translations of the first two phrases provided by the corpus and by the dictionary are the same, but the phrase *postal packet* is an exception. The dictionary translation of *postal packet* is similar to the corpus translation, but not exactly the same. The dictionary translation of the phrase is 小件邮包/*xiao jian you bao* (back translation: *small postal packet*), but the corpus translation is 邮包/*you bao* (back translation: *postal packet*). The dictionary adds the extra meaning of *small* in the translation, but the corpus does not. In this sense, the corpus translation is more accurate than the dictionary translation.

Table 13 contains the remaining 24 A+N phrases which appear as neither subentries nor as examples in the NECD. From the comparison of the six phrases above, it is not surprising that the default translation of these 24 phrases provided by the NECD will not be identical to the translation equivalents that have been found in the corpus.

Table 13: 24 A+N phrases neither as subentries nor as examples

A+N Phrase	Dictionary Default Translation	Corpus Dominant Translation
annual allowance	每年的允许/ <i>mei nian de yun xu</i>	年积金/ <i>nian ji jin</i>
criminal offences	犯罪的冒犯/ <i>fan zui de mao fan</i>	刑事罪行/ <i>xing shi zui xing</i>
final appeal	最后的上诉/ <i>zui hou de shang su</i>	终审/ <i>zhong shen</i>
first column	第一的柱/ <i>di yi de zhu</i>	第 1 栏/ <i>di yi lan</i>
good order	好的次序/ <i>hao de ci xu</i>	良好秩序/ <i>liang hao zhi xu</i> /11
human remains	人的剩余(物)/ <i>ren de sheng yu (wu)</i>	人类遗骸/ <i>ren nei yi hai</i> /41
judicial trustee	司法的受托人/ <i>si fa de shou tuo ren</i>	司法受托人/ <i>si fa shou tuo ren</i>
legal adviser	法律的劝告者/ <i>fa lu de quan gao zhe</i>	法律顾问/ <i>fa lu gu wen</i>
legal officer	法律的官员/ <i>fa lu de guan yuan</i>	律政人员/ <i>lü zheng ren yuan</i>
light bus	轻的公共汽车/ <i>qing de gong gong qi che</i>	小巴/ <i>xiao ba</i> /31
necessary modifications	必要的缓和/ <i>bi yao de huan he</i>	必要的变通/ <i>bi yao de bian tong</i> /20
notifiable workplace	须报告卫生当局的工作场所/ <i>xu bao gao wei sheng dang ju de gong zuo chang suo</i>	应呈报工场/ <i>ying cheng bao gong chang</i>
personal representatives	个人的继承人/ <i>ge ren de ji cheng ren</i>	遗产代理人/ <i>yi chan dai li ren</i>
provisional registration	临时的登记/ <i>lin shi de deng ji</i>	临时注册/ <i>lin shi zhu ce</i>

public bus	公的公共汽车/gong de gong gong qi che	公共巴士/ gong gong ba shi
rateable value	可估价的价值/ke gu jia de jia zhi	应课差餉租值/ ying ke cha xiang zu zhi
reasonable ground	合情合理的地/he qing he li de di	合理的理由/he li de li you/16
registered dentist	已登记的牙医/yi deng ji de ya yi	注册牙医/ zhu ce ya yi
registered scheme	已登记的计划/yi deng ji de ji hua	注册计划/ zhu ce ji hua
residential care	居住的小心/ju zhu de xiao xin	住宿照顾/zhu su zhao gu
restricted licence	有限的许可/you xian de xu ke	有限制牌照/ you xian zhi pai zhao
special category	特殊的种类/te shu de zhong lei	特种/ te zhong
straight line	直的线/zhi de xian	直线/ zhi xian
written permission	写下的允许/xie xia de yun xu	书面准许/shu mian zhun xu/17

The first impression that Table 13 gives to a translator is that the corpus translations are more idiomatic and professional than the dictionary translations. The translation equivalence from the corpus comes nearer to professional terminology than the dictionary default translation does. For instance, *adviser* in *legal adviser* is more like a translation in a legal document when it is rendered as 顾问/*gu wen*; *final appeal* is more like a terminology when translated as 终审/*zhong shen*; and *remains* in *human remains* is more formal when translated as 遗骸/*yi hai*. The corpus translation here may be judged as being better than the dictionary default translation. Another difference between the two column equivalents is that the Chinese adjective suffix “的/*de*” provided by the dictionary is always omitted in the corpus translations. For instance, 直的线/*zhi de xian* of *straight line* has been omitted as 直线/*zhi xian*. Without the suffix “的/*de*”, the corpus translation equivalents are more concise and sound more natural.

For some phrases, there are no translation equivalents for one of the composite elements in the NCED that would let us arrive at the corpus translation. The word *bus* in *public bus* and *light bus* is an example of this, as mentioned before. Another example is *licence* in *restricted*

licence. The corpus translation 牌照/*pai zhao* of *licence* is a Chinese dialect word, used mainly in Southern China, including Hong Kong, but not in the Northern Mainland China. That is why the NCED does not contain this option in it.

It is worth noticing that for some of the above phrases, the dictionary provides similar but not identical phrases. For example, *public transport* (公共交通/*gong gong jiao tong*) is given in the NECD, which is a superordinate of *public bus*; *personal secretary* (私人秘书/*si ren mi shu*) in the NECD is similar to the *personal representatives* of the corpus. However, we cannot derive the same translation equivalent that we find in the corpus from this kind of analogy. By analogy, *public* is 公共/*gong gong*, but the NECD does not provide the translation of *bus* as 巴士/*ba shi* (巴士/*ba shi* is a Chinese word used in Hong Kong, but seldom used on the mainland before 2000); hence, we can never generate the corpus equivalent of 公共巴士/*gong gong ba shi* through using the NECD only. Similarly, *personal representatives* can never be translated as 遗产代理人/*yi chan dai li ren*. If we make an analogy from the NECD, the translation result will be 私人继承人/*si ren ji cheng ren* (= private inheritor), which is far from the original meaning.

Apart from the above mentioned examples, it is impossible to produce the same translation equivalents for the remaining phrases (*final appeal*, *annual allowance*, *legal officer*, *notifiable workplace*, *rateable value* and *residential care*) by referring only to the NECD.

The corpus indicates that *annual allowance* (年积金/*nian ji jin*) is a term, but the entry *allowance* has no translation equivalent for 积金/*ji jin* in the NECD. This is because at the time the dictionary was being compiled, Mainland China had not introduced an allowance system. This fact proves that dictionaries are, to some degree, out-of-date, even when they are newly published.

The above comparison has shown a few of the advantages of the parallel corpus and the limitations of the dictionary when generating the translation equivalence of phrases. Firstly, the dictionary often cannot provide the correct translation equivalents that are found in the parallel corpus. Secondly, due to the length of time required to compile and publish a dictionary, traditional dictionaries may not reflect the current state of the target language. The dictionary always falls behind the rapidly developing language. Thirdly, the traditional dictionary does not include all variations of a translation equivalent used by different people in different areas of a language community.

Therefore, if one takes the corpus translation as a standard rendering, the dictionary gives only three A+N phrases which are exactly the same as the translation provided by the corpus. This means that 90% (27 out of 30) A+N phrases cannot be correctly translated by consulting the general dictionary alone. This explains why translators cannot usually translate correctly into a non-native language by consulting the dictionary.

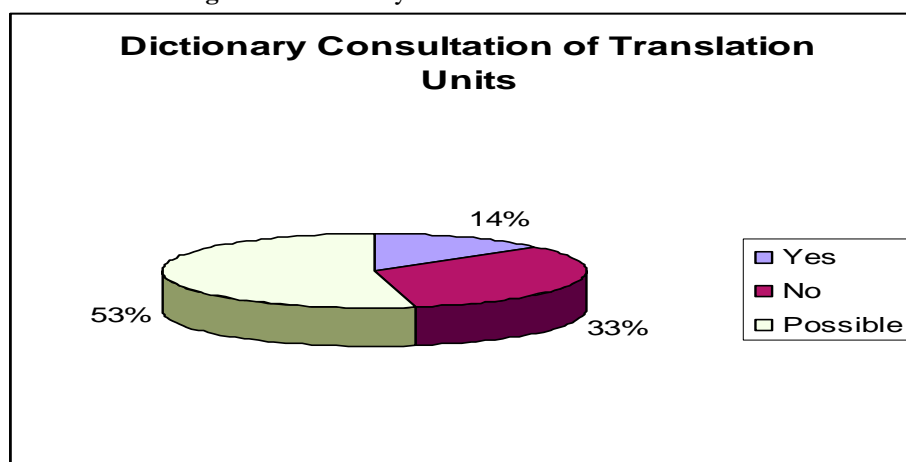
Another comparison is made between the translation equivalents of all the 44 translation units yielded in Chapter 5 and those best combinations of the dictionary translation options. If all

the 44 translation units produced from the 30 phrases are translated by only consulting this dictionary, there are three possible results:

1. The corpus translation equivalents can be directly found in the dictionary;
2. There is no way to achieve the corpus translation from the dictionary;
3. There is a possibility of achieving the corpus translation but it will depend on how the translator selects the translation options for each word, and how these will be combined.

The result is shown in the following graph. Only seven translation units yielded the same translation equivalents as the corpus ones (marked as “Yes”); 17 translation units did not yield the same translation equivalents at all (marked as “No”); the remaining 20 have the possibility of yielding the same translation equivalents but it all depends on how careful the translators are in selecting and combining the dictionary translation options (marked as “Possible”). In the best situation, all the 20 “Possible” cases have been translated in the same way as the corpus translation, and this is called the best combination. Then, there are still 33% translation units which cannot be correctly translated by consulting the dictionary, as shown in Figure 3.

Figure 3: Dictionary consultation of translation units



6.3 A Comparison with the English-Chinese Glossary of Legal Terms

The comparison of Section 6.3 has demonstrated that when it comes to translating terminology, the ordinary general dictionary cannot be of much help. It thus raises an interesting question: how much help can a specific dictionary provide? In this section, a dictionary of legal terminology will be used and the results will be compared with the corpus translation.

Three kinds of results are obtained when searching for the 30 phrases in the ECGLT:

- (1) Phrases not found in the ECGLT (8 cases);
- (2) Parts of phrases found in the ECGLT, but the translation equivalents in the ECGLT differ little from the results in the HKLDC (4 cases);
- (3) Phrases appear in the ECGLT, and the translation equivalents in the ECGLT are the same as in the HKLDC (18 cases);

(1) Eight A+N phrases are not found in the ECGLT, they are:

Straight line, human remains, public bus, light bus, postal packet, special category, registered scheme, and internal combustion

As mentioned before, these phrases occur approximately 100 times in the HKLDC; in other words, they are very common in legal documents. The reason why they are ignored may be either that they are too common to be entered into the dictionary (e.g. *public bus* and *light bus*) or they are not legal terms (e.g. *internal combustion engine* may be regarded as a technical

term). Whatever the reason why they have been ignored, it is not helpful for the non-native speaker who is attempting a translation. If neither the specialized dictionary nor the general dictionary has listed the phrase (e.g. *light bus*), it will be a problem for the non-native translator.

Strictly speaking, the phrase *registered scheme* is not found in the ECGLT. The ECGLT lists many collocations of *registered* and *scheme*, but it does not list them as a combined phrase. In the ECGLT, *registered* has three translation equivalents when it forms other units: 登记/*deng ji*, 注册/*zhu ce*, and 挂号/*gua hao*; *scheme* is always translated as 计划/*ji hua* in all its larger translation units. The ECGLT thus asks the translator to deduce the equivalent of *registered scheme*; should it be 登记计划/*deng ji ji hua*, 注册计划/*zhu ce ji hua*, or 挂号计划/*gua hao ji hua*? If a translator does not have adequate knowledge of Chinese, s/he may arrive at a non-idiomatic translation.

(2) Four phrases are partly found in the ECGLT, they are:

long term, conclusive evidence, personal representatives, legal adviser and registered scheme.

The term *long term* is located under the headword *business* as *long term business* 长期业务/*chang qi ye wu*. But another translation unit of *long term* — *long term interest* (see Chapter 5) is ignored. This shows a deficiency of the ECGLT. It does not list all relevant translation units.

The term *conclusive evidence* can be derived both from the two headwords *evidence* and *conclusive*. However, the ECGLT gives only one of its translation equivalents—*不可推翻的证据/bu ke tui fan de zheng ju*. Another translation equivalent, *确证/que zheng*, is ignored. According to the BLIS database, to which the ECGLT is linked, there are 25 sections which have *不可推翻的证据/bu ke tui fan de zheng ju* as translation equivalents and another 25 sections have *确证/que zheng* as its translation equivalent. The two translation equivalents have about the same frequency and should therefore not be ignored. This is clear evidence that the ECGLT is not corpus-based and that it therefore does not give complete coverage of translation equivalents.

There is no term *personal representatives* found in the ECGLT but a similar phrase *legal personal representative* is listed under the headword *personal*. According to the sampling study, *personal representatives* can always be used alone as an independent translation unit without the word *legal* before it. Since Chinese does not make the difference between the singular and plural forms, the reader may regard the phrase *personal representatives* as a part of *legal personal representative*. Although the reader can use analogy to identify the translation equivalent of *personal representatives*, this proves that the ECGLT has not correctly identified all the translation units from the database to which it is linked.

Moreover, the translation equivalent of *legal personal representative* that the ECGLT provides is a little misleading. The corpus dominant translation equivalent is sometimes not listed but, instead, the much less frequent one is found. The corpus dominant equivalent is *合法遗产代理人/he fa yi chan dai li ren* instead of *法定遗产代理人/fa ding yi chan dai li ren*

provided by the ECGLT. The BLIS database also shows that 法定遗产代理人/*fa ding yi chan dai li ren*, only appears in three sections of legal documents issued between 2000 and 2002, while the equivalent found in the corpus, 合法遗产代理人/*he fa yi chan dai li ren*, is used more often (in 41 sections issued between 1997 and 2004 in the BLIS database). Why the ECGLT does not list the dominant translation equivalent, but the much less frequent one, is a puzzle.

The problem of *legal adviser* is similar to that of *personal representatives*. The phrase *legal adviser* is listed neither under *legal* nor *adviser*, but under the unexpected headword *professional*. There is no *legal adviser* standing alone as a whole composite phrase, but only *professional legal adviser*. This will not only make it difficult for the user to find *legal adviser*, especially in a printed dictionary, but could also lead the user to draw the wrong conclusion for the translation by analogy from the longer phrase *professional legal adviser*.

(3) The remaining 18 A+N phrases can be found in the ECGLT:

legal officer, residential care, criminal offences, annual allowance, written permission, first column, notifiable workplace, listed company, registered dentist, good order, provisional registration, judicial trustee, final Appeal, necessary modifications, rateable value, restricted licence, reasonable ground, and medical officer.

Each phrase is listed under a headword which is part of the phrase. However, sometimes they are listed under both components (e.g. *annual allowance* can be found both under *annual* and *allowance*), and sometimes, under only one (e.g. *first column* under *column*).

One exception is *final appeal*. Like the corpus sample study, *final appeal* is only part of the larger unit of *(the) court of final appeal*. The ECGLT, however, puts the phrase under *court*, but not under *final* or *appeal*. This is an inconsistency often found in bilingual dictionaries and there is no standard solution. Should these units be listed under all their components, or under only one? If they are listed under all their components, this takes up space; if they are listed under only one headword, then the user spends more time looking it up.

The ECGLT provides correct translation equivalents for the majority of the 30 A+N phrases and is very much better than a general language dictionary. Furthermore, the online version of ECGLT is linked to the bilingual law database, which greatly improves accessibility. However, there are still 40% of the phrases (12 out of 30) which cannot be found in the ECGLT. This means that there is only a 60% translation accuracy if the user consults the ECGLT only.

The reason why the three different outcomes were found is: the ECGLT is not completely corpus-based. Our analysis has also established the following facts:

- (1) 27% of the 30 A+N phrases cannot be found in the ECGLT at all.
- (2) Although some of them can be found in the collocation column, the reader cannot identify under which headword they are listed. This increases the difficulty for users of the printed versions. Sometimes the same phrase is repeatedly provided under several headwords; sometimes it is listed only under one headword, sometimes even under an unlikely headword which the user cannot predict.

- (3) The ECGLT sometimes fails to provide the dominant equivalent, as shown in the HKLDC.
- (4) The same headword is listed several times, only to present various composite phrases including it; sometimes the ECGLT provides too many equivalents for the same translation unit.

6.4 Conclusion

In this chapter, through comparisons of the general dictionary (the NECD) and the specialised legal glossary (the ECGLT), some advantages of corpus-based lexicography over traditional dictionaries have been illustrated. The parallel corpus provides accurate and natural translation equivalents, while traditional dictionaries often do not. Also, traditional dictionaries can fall behind the rapidly developing language, while translation databases using incremental parallel corpora perform better in this respect. Finally, traditional dictionaries do not include all variations used in different areas of a language community. These deficiencies can be remedied if future bilingual dictionaries are based on bilingual corpora.

Chapter 7

Conclusion

The translation unit is an important research topic in parallel corpus study. This dissertation seeks a way to define and describe the translation unit in the English-Chinese bilingual context by using the corpus linguistic theory of a unit of meaning, a concept which originally was used in a monolingual context. As an aid to the reader, the first section of this final chapter (Section 7.1) restates the research problem and reviews the major methods used in the study; Section 7.2 summarizes the main results obtained in the previous thesis; Section 7.3 discusses the meaning of this study and points out further work which needs to be done in the future.

7.1 Research Problems and Methods

Traditional bilingual dictionaries are deficient in important ways. Their main disadvantage is that they do not offer enough support to translators wishing to translate into a target language which is not their native language. This deficiency is inherent, deriving from the practice of treating single words as their standard lemma. Since most single words are polysemous, traditional dictionaries provide several translation options for each word; the problem of ambiguity in translation thus arises. These translation options may be helpful to those who have a large vocabulary in the target language or who know which option to choose in different cases, but may not be helpful to those who have a more limited vocabulary in the target language. This dissertation argues that parallel corpora are an essential supplement to

traditional dictionaries because the translation units and their translation equivalents which largely exist in parallel corpora, provide a solution to the problem of ambiguity.

The translation unit, in this dissertation, is the basic monosemous unit in translation (Teubert, 2005). It is an unambiguous unit that has only one meaning and therefore is assumed to have only one translation equivalent in the target language. In other words, the translation unit is a kind of unit of meaning but from the target language perspective. Like the notion of unit of meaning in corpus linguistics, the concept of translation units comes from the text combination principle of collocation: “the choice of one word conditions the choice of the next and of the next again” (Sinclair, 1994:22; 2004:19). Since the unit of meaning in corpus linguistics is regarded as much more extensive and varied than a single word (Sinclair, 2004: 39-40), the translation unit is assumed mostly to be larger than a single word in source texts. Translation units and their target equivalents are the technical media through which the parallel corpora can be used and reused to benefit the future translations. The criteria for identifying the translation units are proposed in this dissertation (Chapter 2).

Apart from defining and describing the concept of translation unit, this dissertation has also tried to show what the translation units look like in a 10 million words/characters English-Chinese specialised parallel corpus, the Hong Kong Legal Document Corpus (HKLDC) (See Chapter 3). The extraction started from the A+N pattern. This pattern was chosen because examples are most likely to be translation units, as well as, because they are one of the most common and stable lexical patterns so that we can yield enough occurrences of them and their equivalents (see Chapter 4). Among all the A+N patterns extracted, 30 sample phrases were

selected and analysed. It was found that not all the 30 phrases are automatically translation units; some have to be extended (see Chapter 5).

To test the idea that parallel corpora can be the necessary supplement to traditional bilingual dictionaries, the extracted translation units and the translation equivalents were compared to those from the traditional dictionaries. Two bilingual dictionaries — the NECD and ECGLT (see Chapter 6) were chosen for the comparisons. One is a general dictionary while the other is a special legal glossary. The comparison focused on two sides: one is on whether the translation units can be yielded from the dictionaries or not; the other on whether the translation equivalents can be yielded or not. Two comparisons are made with the general dictionary. The first one compares the equivalents of the 30 A+N phrases to the default translations yielded from the dictionary; the second one compares the equivalents of the 44 translation units compared to the best dictionary translation (Chapter 6).

7.2. Main Results

The results of this study strongly suggest that when using a parallel corpus it is possible to identify translation units as the smallest monosemous units, and that these units are often larger than single words. Through the pilot study of the 30 A+N phrases and their translation equivalents, the hypothesis that one complete translation unit has one translation equivalent in an ideal situation has preliminarily been shown to be correct in a specialised corpus like the HKLDC. That is, each translation unit candidate has only one translation equivalent. Otherwise the translation equivalents are either synonymous, which means they can replace each other, or else the candidate is part of different translation units. If a translation unit

candidate belongs to the latter case, more text should be added until monosemous translation equivalents of complete translation units are yielded. It was found that 13 of the 30 A+N phrases were complete translation units while the rest were not. Those that were not, were extended to complete translation units, until each of these translation units had one translation equivalent in Chinese. If, for any reason, they have more than one translation equivalent, these equivalents are either synonymous, like 小巴/*xiao ba* and 小型巴士/*xiao xing ba shi* (which can replace each other for *light bus*) or they belong to several translation units. In this case, more text needs to be added until each translation unit has only one translation equivalent. The process of the extraction is an application of the equation: Translation unit = Node + Context. The node here is the A+N phrase. The context of a node is the words or other factors which make the node monosemous.

It is worth pointing out that the context of a node may not just be the adjacent words to the left or right, but may include some space further from the node. The context can be specific (e.g. words) as well as abstract (e.g. the whole domain such as in *medical officer* (Chapter 5)). The difficulty of the abstract context is that it is not easy for a computer to recognize it. In the 30 A+N phrases explored in this thesis, the most variable and the most complicated expansion of translation unit is *good order* (see Chapter 5).

Another result found in the study is that the translation unit and translation equivalents yielded from the parallel corpus are better than the combinations of single words from traditional dictionaries. Using the general dictionary NECD, two comparisons were made. The first, is that the translation equivalents of the 30 A+N phrases yielded from the parallel

corpus were compared to those default translation equivalents; the second, is that the corpus translation equivalents of the 44 translation units were compared to those best combinations in the general dictionary NECD. The default translation equivalent means the combination of the first translation option of each word provided by the dictionary in a translation unit or a phrase. The best combination means the combination of each word in the translation unit or a phrase which is closest to the translation equivalents provided by the corpus. This traditional bilingual dictionary could yield only 3 default translation equivalents which were the same as those from the parallel corpus; for the remaining 90% of the phrases, their equivalents were either not found in the dictionary or were different from what the corpus provided. In the best combination situation, there were still 33% of translation units which could not be correctly translated by consulting this dictionary. The situation in special legal glossary ECGLT was better, but still 40% of corpus equivalents could not be yielded by consulting it. Parallel corpora provide correct and natural translation equivalents, while the dictionaries often do not. These comparisons have provided direct evidence that parallel corpora have advantages in corpus-based lexicography over the traditional bilingual dictionaries, and provide an alternative way to improve conventional dictionaries.

7.3. Relevance of this Study and Further Works

This study applies the model of the translation unit to bilingual parallel corpora study and provides a possible way of exploring translation units and translation equivalents. It has shown that useful language patterning, such as A+N, can be drawn from the parallel corpus. Based on this pattern, translation units can be extracted according to their translation equivalents.

The translation unit is a useful concept developed from the unit of meaning in corpus linguistics and by using this concept it is possible to reduce some problems of conventional word-based ambiguity in translation and Machine Translation. If the hypothesis is generally true in all circumstances, that is, if each translation unit has only one translation equivalent, and the translation is consistent enough, the translation unit will be very meaningful in disambiguation. If the unambiguous translation unit replaces the single word as the standard lemma in a dictionary, it will help not only those translators with less knowledge in the target language, but will also help Machine Translation.

Translation units also provide a way of locating the boundary of a unit of meaning in the original text through the view of another language. Experienced translators will be aware of where the boundary of a translation unit is; in other words, they will know where the boundary of a unit of meaning is. The monosemous translation equivalent in the target language can help us to determine the unit of meaning in the original language. It can help the source language learner to learn and can guide our teachers as well. For instance, *special category*, *final appeal* and *restricted licence* can be respectively expanded to *special category space(s)*, *(the) court of final appeal* and *restricted licence bank*.

Moreover, as the repository of translation units and translation equivalents, parallel corpora have huge research potential, especially to bilingual or multilingual lexicography. They can be used to form or improve lexica, translation bases, word banks or bilingual dictionaries. They are an indispensable supplement of translation aids. Parallel corpora contain the

experienced work of translators, and therefore the translation equivalents of some translation units in parallel corpora are more accurate and natural than the direct combination of dictionary translation options. For instance, there is no way to translate *residential care home* as 安老院/*an lao yuan* just by consulting a dictionary. By consulting parallel corpora, translators will find more idiomatic, ready-to-use translation equivalents; consequently, the whole text will be better translated. Once translation units have replaced single words as the standard lemma in dictionaries, the consultation process will also largely be shortened.

Nevertheless, more work needs to be done in this pioneering area. First, the methodology and results need to be tested using a larger scale general corpus. This dissertation could only analyse 30 typical A+N phrases based on the specialised parallel corpus HKLDC. The legal document belongs to LSP (Language for Specialised Purpose) which has some characters that a general corpus may not have. It would be interesting to see whether the same results would be found with general texts. Therefore, a large scale general corpus is needed to test the methodology and results. In addition, we should bear in mind that due to different cultures and histories, different languages are composed of a number of different strata. The more strata a language is composed of, the more colourful it is; therefore, more synonymous translation equivalents may appear in a large scale general corpus.

Second, the hypothesis needs to be tested with other patterns such as verbal patterns and phrases. As part of this study, the N+N (Noun +Noun) pattern was also extracted from the HKLDC, but due to the limitation of space, only the 30 A+N phrases and their translation equivalents are discussed in this thesis.

Lastly, software should be developed to identify translation units in order to explore parallel corpora fully and more automatically. This study highlights the fact that searching for translation units is readily computable if each translation unit has only one translation equivalent. Such a result should help with the design of auxiliary software to search for translation units and their equivalents.

References

- A Chinese-English Dictionary*. 1993. The Commercial Press. Beijing.
- A New English-Chinese Dictionary*. (century ed.). 2000. Shanghai Translation Publishing House.
- Aitchison, J. 1987. *Words in the Mind: An introduction to the Mental Lexicon*. Oxford: Blackwell.
- Aramaki, E., Kurohashi, S., Sato, S. and Watanabe, H. 2001. Finding Translation Correspondences from Parallel Parsed Corpus for Example-Based Translation. *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain, pp27-32, 2001.
- Baker, M. 1992. *In Other Words: a course book on translation*. London/New York: Routledge.
- Baker, M. 1993. Corpus Linguistics and Translation Studies: Implications and Applications". In M.Baker, G. Francis and E. Tognini-Bonelli (Eds). *Text and Technology: in Honour of John Sinclair*. Amsterdam and Philadelphia: John Benjamins, pp. 233-250.
- Baker, M. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* 7 (2): 223-243.
- Baker, M. 1996. Corpus-Based Translation Studies: the challenges that lie ahead. In H. Somers, (ed.) *Terminology, LSP and Translation Studies in Language Engineering in honour of Juan C. Sager*. Pp175-186. Amsterdam: Benjamins.
- Barkhudarove, L. 1993. The Problem of the Unit of Translation. In Palma Zlateva (Ed.). *Translation as Social Actif*. London and New York: Routledge.
- Biber, D. et al. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bowker, Lynn. 1998. Using Specialized Monolingual Native-language Corpora as a translation Resource: A Pilot Study.
- Burkhanov, I. 2003. Pragmatic Specifications: Usage indications, labels, examples; dictionaries of style, dictionaries of collocations. In P.V. Sterkenburg (ed.), *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company. Pp 102-113.
- Chan, Sin-wai. 2002. The Making of TransRecipe: A Translational Approach to the Machine Translation of Chinese Cookbooks. In S. Chan (ed). *Translation and Information Technology*. Hong Kong: The Chinese University Press, pp.1-20.

- Chang, B., Danielsson, P. and Teubert, W. 2005. Chinese-English Translation Database: Extracting Units of Translation from Parallel Texts. In G. Barnbrook, P. Danielsson and M. Mahlberg (eds). *Meaningful Text: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. London/New York: Continuum. Pp131-142.
- Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Danielsson, P. and Ridings, D.1997. *Practical Presentation of a "Vanilla" Aligner*. Available at : <http://nl.ijs.si/telri/Vanilla/doc/ljubljana/>
- Danielsson, P. and Ridings, D. 1999. Corpus and terminology: software for the translation program at Göteborgs Universitet. In *Multilingual Corpora in Teaching and Research*. S. P. Botley et al (eds). Amsterdam: Rodopi.
- Ebeling, J. 1998. Contrastive Linguistics, Translation and Parallel Corpora. *Meta XLIII* 4.
- Firth, J. R. 1935. The Technique of Semantics. *Transactions of the Philological Society*. Pp 36-72.
- Firth, J. R. 1957a. A Synopsis of Linguistics Theory 1930-1955. In F.R.Palmer (ed), *Selected Papers of J.R.Firth, 1952-1959*. London: Longman.
- Firth, J. R. 1957b. *Papers in Linguistics 1934 – 1951*. London: Oxford University Press.
- Francis, G. 1993. A Corpus Driven approach to grammar: Principles, methods and examples. In M. Baker, G. Francis and E. Tognini-Bonellie (eds), *Text and Technology*. Amsterdam: Benjamins, pp.137-56.
- Gale, W. A. and Church, K.W. 1993. A Program For Aligning Sentences in Bilingual Corpora. *Computational Linguistics*. 19(1): 75-102.
- Geeraerts, D. 2003. Meaning and Definition. In P.V. Sterkenburg (ed.), *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company. Pp 83-101.
- Hunston, S. and Francis, G. 2000. *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins Publishing Company.
- Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Kennedy, G. (2000). *An Introduction to Corpus Linguistics*. Beijing: Foreign Language Teaching and Research Press.

- Kenny, D. (2001). *Lexis and Creativity in Translation. A Corpus-based Study* Manchester: St. Jerome.
- Kit, C., Webster, J. J., Sin, K.K., Pan, H., and Li, H. 2004. Clause Alignment for Hong Kong Legal Texts: A lexical-based approach. In *International Journal of Corpus Linguistics*. Vol.9:1(2004), pp 29-51.
- Laviosa-Braithwaite, S. 1996: Comparable Corpora: towards a corpus linguistics methodology for the empirical study of translation. In M.Thelen, and B. Lewandowska-Tomaszczyk (eds). *Translation and Meaning*. Maastricht: UPM. Part 3, pp. 153-163.
- Lyons, J. 1968. *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.
- Lyons, J. 1977. *Semantics*. Cambridge: Cambridge University Press.
- Malmkjær, K. 1998. Unit of Translation. In M. Baker (ed.) *Routledge Encyclopedia of Translation Studies*. London and New York: Routledge. Pp 286-288.
- Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-based Approach*, Oxford : Clarendon Press.
- Palmer, F.R. 1968. *Selected Papers of J.R.Firth 1952-59*. London: Longman.
- Ramm, W. 2004. Sentence boundary adjustments in Norwegian-German and German-Norwegian translations: First results of a corpus-based study. In K. Aijmer and H. Hasselgård (eds.) *Translation and Corpora: Selected Papers from the Göteborg-Oslo Symposium 18-19 October 2003*. Göteborg: Acta Universitatis Gothoburgensis.
- Ribeiro, A., Lopes, G. and Mexia, J. (2001) "Extracting Translation Equivalents from Portuguese-Chinese Parallel Texts". In Lee Sangsup (ed.), *Proceedings of Asialex 2001: Asian Bilingualism and the Dictionary — The Second International Congress of the Asian Association for Lexicography (Asialex)* (Seoul, South Korea, 2001 August 8-10), pp. 225-230.
- Sato, S. and Nagao, M. 1990. Toward Memory-based Translation. In *Proceedings of Thirteenth International Conference on Computational Linguistics (COLING-90)*. Helsinki, Finland. Pp 247—252.
- Schmied, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. September. Available at <http://www.ims.uni-stuttgart.de/~schmid>
- Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP

- Sinclair, J. M. 1994. Trust the Text. In Malcolm Coulthard (ed.), *Advances in Written Text Analysis*. London/New York: Routledge.
- Sinclair, J. M. 1996. The Search for Units of Meaning. *Textus IX*. pp 75-106.
- Sinclair, J. M. 1998. The Lexical Item. In Weigand, E. (ed.) *Contrastive Lexical Semantics*. Amsterdam: John Benjamins, 1-24.
- Sinclair, J. M. 2001. Review of the Longman Grammar of Written and Spoken English. *International Journal of Corpus Linguistics*. 6 (2), 339-359.
- Sinclair, J. M. 2003. 'Corpora for lexicography'. In van Sterkenburg, Piet (ed.) *A practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company.
- Sinclair, J. M. 2004. (Edited with R. Carter). *Trust the Text: Language, corpus and discourse*. London/New York: Routledge.
- Stubbs, M. 1993. British Tradition in Text Analysis: From Firth to Sinclair. In Baker et al. (eds.) *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins Publishing Co. pp 1-33.
- Stubbs, M. 2002. *Words and Phrases: Corpus Study of Lexical Semantics*. Oxford: Blackwell Publishing.
- Swanepoel, P. (2003). 'Dictionary typologies: A pragmatic approach'. In van Sterkenburg, Piet (ed.) *A practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company.
- Sun, I. 2003. Chinese-English *TranslationBase* Project Work Report (1-5). Powerpoint slides. Presented at the Centre for Corpus Linguistics, University of Birmingham.
- Teubert, W. 1996. Comparable or Parallel Corpora? *International Journal of Lexicography*. Vol. 9, No. 3. Oxford University Press. Pp 238-264.
- Teubert, W. 1999. |Starting with *Trauer*. Approaches to Multilingual Lexical Semantics. In F. Kiefer et al (eds). *Papers in Computational Lexicography Complex'99*. Budapest: Linguistics Institute Hungarian Academy of Science.
- Teubert, W. 2001. Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics*. Vol. 6 (Special issue). Pp125-153.
- Teubert, W. 2002. The Role of Parallel Corpora in Translation and Multilingual Lexicography. In B. Altenberg and S. Granger (eds.). *Lexis In Contrast*. Amsterdam / Philadelphia : Benjamins. Pp 189 – 214.

- Teubert, W. 2003. Collocations, Parallel Corpora and Language teaching. In: *Selected Papers from the Twelfth International Symposium on English Teaching*. Taipei: English Teacher's Association, 143-156.
- Teubert, W. 2004. Language and Corpus Linguistics. In M.A.K. Halliday et al. *Lexicology and Corpus Linguistics: An Introduction*. London/New York: Continuum. Pp73-112.
- Teubert, W. 2005. My Version of Corpus Linguistics. *International Journal of Corpus Linguistics*. 10:1, 1-13.
- Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.
- Tufis, D. 2001. Computational bilingual lexicography:automatic extraction of translation dictionaries. In *Journal of Information Science and Technology, Romanian Academy* 4(3).
- Webster, J. J, Sin, K. K. and Hu, Q. 2002. The Application of Semantic Web Technology for Example-based Machine Translation (EBMT). In S.W Chan (ed.) *Translation and Information Technology*. Hong Kong: The Chinese University Press.
- Wright, S.E. 1997. Term Selection: The Initial Phrase of Terminology Management. In S. E. Wright and G. Budin (eds.). *Handbook of Terminology Management* (Vol.1): *Basic Aspects of Terminology Management*. Amsterdam: John Benjamins Publishing Company.
- Wu, D. 1995. Grammarless Extraction of Phrasal Translation Examples from Parallel Texts. *TMI-95, Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*. Leuven, Belgium. 1995 July, v2, 354-372. Available at <http://www.cs.ust.hk/~dekai/>
- Yen, Tony Yuan-Ho. 2001. Hong Kong's Bilingual Laws Programme. In Sin-wai Chan (ed.) *Translation Hong Kong: Past, Present and Future*. Hong Kong: The Chinese University Press. 249-254.
- Yamamoto, K., Matsumoto, Y. and Kitamura, M. 2001. A Comparative Study on Translation Units for Bilingual Lexicon Extraction. *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*. pp.87-95, Toulouse, July 2001
- Zgusta, Ladislav. 1984. Translational Equivalence in the Bilingual Dictionary. In R.R.K.Hartmann (ed.) *LEXeter's 83 Proceedings: papers from the International Conference on Lexicography at Exeter, 9 – 12 September 1983*. Tübingen : Max Niemeyer. pp147-154.

Software

ConcApp : http://vlc.polyu.edu.hk/pub/concapp/
ParaConc : http://www.ruf.rice.edu/~barlow/parac.html , First BETA version, 1996
TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Appendices

Appendix 1: Email from the Department of Justice

From: blis@doj.gov.hk
Subject: RE: enquiry about BLIS-

Dear Ms Wang,

Mrs Tammy Fung has forwarded to me your enquiry on the process of our bilingual legislative drafting after 1st July 1997.

2. You may be aware that both the English and the Chinese languages are the official languages of Hong Kong. As the 2 languages possess equal status, the English text and the Chinese text of a piece of legislation in Hong Kong are equally authentic, that is, they are authentic versions of the law which the court can look at in order to ascertain the law. Out of this consideration, legislative drafting at the Law Drafting Division is entrusted with officers of the Government Counsel Grade ("GC Grade"), who must be solicitors or barristers. Counsel working at the Law Drafting Division have to be proficient in both of the official languages.

3. As far as the drafting process is concerned, the preparation of the Chinese texts of legislation is essentially a translation process. The English text is normally drafted first, and then rendered into Chinese. This is particularly true when the English text is drafted by anglophone counsel who cannot write Chinese. Bilingual draftsmen also prefer to do so as they have studied law in English, which is the working tool of the common law.

4. Occasionally, in cases where the draftsman drafts both the English text and the Chinese text of a piece of legislation, he may prefer to draft part of the text or a particular provision in the Chinese language first and then render it into the English language. Such deviation from the practice displays occasional personal preferences rather than the norm. On the other hand, a draftsman responsible for the bilingual drafting of a piece of legislation always has the convenience of modifying the language of any provision of the English draft to align with the corresponding Chinese draft developed from the initial Chinese translation of the provision.

5. I hope you find the above information useful. Please feel free to contact us again if you have any other questions on the use of BLIS.

Yours sincerely,

Mabel Cheung
Government Counsel
Law Drafting Division
Department of Justice
Government of the HKSAR

Appendix 2: Perl 1

- (1) Open the English text in the corpus;
- (2) Read a sentence from the English texts;
- (3) Read a word from the sentence;
- (4) Check whether the word is the first word of the A+N phrase.
- (5) If yes, then check whether the next word is the second word of the A+N phrase. If yes,
a context sentence has been found and so the sentence is saved into a file (File 2);
- (6) Repeat step (3) - (5), until the end of the sentence has been reached;
- (7) Repeat step (2) - (6), until the end of the English texts in the corpus has been reached.

Appendix 3: Perl 2

- (1) Open the file that stores the extracted English sentences, i.e. File 2;
- (2) Read each extracted English sentence in File 2;
- (3) Extract the first word from this sentence, set the word to be its sentence ID number and save it to another file (File 3);
- (4) Repeat steps (2)-(3), until the end of file 2 has been reached.

Appendix 4: Perl 3

- (1) Open File 3 which stores the English sentence ID number;
- (2) Read a sentence ID number from the file;
- (3) Open the Chinese texts in the corpus;
- (4) Read a sentence from the Chinese texts;
- (5) Extract the first word from this sentence, and set the first word to be its ID number;
- (6) Check whether the Chinese sentence ID number equals to the English sentence ID number coming from the file;
- (7) If yes, then a corresponding Chinese sentence with the same sentence ID number as the English sentence has been found, so this Chinese sentence is saved into a file (File 4)., Otherwise, repeat step (3) – (6), until arrive at the end of the Chinese texts;
- (8) Repeat step (2) - (7), until the end of File 3 has been reached.

Appendix 5 : Perl 4

Given an English A+N phrase,

- (1) Open the English texts in the corpus;
- (2) Read a sentence from the English texts;
- (3) If this sentence contains the phrase, then save the sentence to a file called English sentence file; Extract the first word of this sentence as the sentence ID number;
- (4) Open the Chinese texts in the corpus;
- (5) Read a sentence from Chinese corpus;
- (6) Check whether the Chinese sentence has the same sentence ID number as the English sentence ID number;
- (7) If yes, the corresponding Chinese sentence has been found. Manually check the English-Chinese sentence pair; and extract the translation equivalent; if it is a new translation equivalent different from existing ones, then save it into a file called translation unit file.
Go to step (9).
- (8) Repeat step (2) – (7), until the end of Chinese texts has been reached.
- (9) Repeat step (2) – (8), until the end of English texts has been reached.

Appendix 6:10 Sample Extraction of *conclusive evidence* and its Chinese Equivalents

JJ+NN = conclusive evidence;

Chinese equivalents of *conclusive evidence*: 确证(6) and 不可推翻的证据(4)

441 A certificate issued by the Chief Secretary for Administration that any property vested in a public officer immediately before a resolution under this section takes effect has been transferred by virtue of the resolution to another public officer shall be **conclusive evidence** of the transfer.

441 如政务司司长发出证明书，证明紧接在根据本条通过的决议生效前赋给某一公职人员的财产，已凭借该项决议移转给另一公职人员，该证明书即为该项财产移转的**确证**。

5608 A certificate of the Official Receiver that a person has been appointed trustee under this Ordinance shall be **conclusive evidence** of his appointment.

5608 由破产管理署署长发出以证明某人已根据本条例获委任为受托人的证明书，即为该人获该项委任的**确证**。

8372 Nothing in this section shall prejudice the operation of any enactment whereby a finding of fact in any matrimonial proceedings is for the purposes of any other proceedings made **conclusive evidence** of any fact.

8372 如藉任何成文法则的实施，任何婚姻法律程序中的任何事实的裁断，就任何其它法律程序而言，成为任何事实的**不可推翻的证据**，则本条的规定不得损害上述成文法则的实施。

8375 In an action for libel or slander in which the question whether a person did or did not commit a criminal offence is relevant to an issue arising in the action, proof that, at the time when that issue falls to be determined, that person stands convicted of that offence shall be **conclusive evidence** that he committed that offence; and his conviction thereof shall be admissible in evidence accordingly.

8375 在任何永久形式诽谤或短暂形式诽谤的诉讼中，如某人有否犯某刑事罪行的问题与在该诉讼中出现的争论点有关联，而在对该争论点予以裁定的时间，有证明该人仍就该罪行被定罪，则该证明即为他曾犯该罪行的**不可推翻的证据**，而他就该罪行的定罪亦据此可接纳为证据。

8460 Without prejudice to subsection, a person shall not be compelled by virtue of an order under section 76 to give any evidence if his doing so would be prejudicial to the security of the United Kingdom, Hong Kong, or any other territory for which the United Kingdom is responsible under international law; and a certificate signed by or on behalf of the Chief Secretary to the effect that it would be so prejudicial for that person to do so shall be **conclusive evidence** of that fact.

8460 如某人提供证据会损害联合王国、香港、或任何其它地区的安全，则在不损害第

款的原则下，不得凭借根据第 76 条所作出的命令强迫该人提供证据；而由布政司或他人代其签署、说明该人提供证据会有如此损害的证明书，即为该事实的**不可推翻的证据**。

8537 A certificate purporting to be signed by the Registrar and certifying that any deposition to which such certificate is attached, together with any document or thing exhibited or annexed thereto, has been received by him pursuant to a letter of request issued by him under section 77E in respect of any criminal proceedings referred to in the certificate, shall on its production without further proof be admitted in those criminal proceedings as **conclusive evidence** of the facts contained therein.

8537 任何证明书如看来是由司法常务官签署，并核证该证明书所随附的任何书面供词已由司法常务官依据他根据第 77E 条就该证明书记内所提述的任何刑事法律程序发出的请求书收取，则在该刑事法律程序中一经交出，无须再加证明，即须接纳为该证明书记内所载事实的**不可推翻的证据**。

9768 14. A certificate signed by the Chief Executive of the Corporation that an instrument of the Corporation purporting to be made or issued by or on behalf of the Corporation was so made or issued shall be **conclusive evidence** of that fact.

9768 14. 一份由公司总裁签署的证明书，证明一份看来是由公司或代公司订立或发出的文书是由公司或代公司订立或发出者，即为该事实的**确证**。

13412 A statement, contained in any instrument coming into operation after the commencement of this Ordinance by which a new trustee is appointed for any purpose connected with land, to the effect that a trustee has remained out of Hong Kong for more than 12 months or refuses or is unfit to act, or is incapable of acting, or that he is not entitled to a beneficial interest in the trust property in possession, shall, in favour of a purchaser of a legal estate, be **conclusive evidence** of the matter stated.

13412 任何文书如在本条例生效后实施并为任何与土地有关的目的而委任新受托人，其内如载有任何陈述，意思是受托人不在香港超过 12 个月、拒绝或不适合作为受托人、无行为能力作为受托人或受托人无权享有在管有中的信托财产的任何实益权益，则为了法定产业权买家的利益，该陈述即为所述事项的**确证**。

13499 the fact that the order has been so made shall be **conclusive evidence** of the matter so alleged in any court upon any question as to the validity of the order; but this section does not prevent the court from directing a reconveyance or surrender or the payment of costs occasioned by any such order if improperly obtained.

13499 则在该命令的有效性的问题上，该命令已如此作出此一事实，在任何法院即为如此指称的事项的**确证**；但本条并不阻止法院指示将土地再转易或退回，或指示支付不适当地取得该命令所引致的费用。

14163 A certificate of incorporation issued by the Registrar in respect of any association shall be **conclusive evidence** that all the requirements of this Ordinance in respect of registration and of matters precedent and incidental thereto have been complied with, and that the association is a company authorized to be registered and duly registered under this Ordinance.

14163 处长就任何组织所发出的公司注册证书，即为以下事项的**确证**：本条例中与注册

有关的所有规定及与注册的先决及附带事宜有关的所有规定已获遵从，以及该组织是一间根据本条例获批准注册并已妥为注册的公司。